SAS-, SPSS- und STATA-Programme zur Berechnung der Varianz von Populationsschätzern im Mikrozensus ab 1996

Schimpl-Neimanns, Bernhard; Rendtel, Ulrich

Veröffentlichungsversion / Published Version
Arbeitspapier / list

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
GESIS - Leibniz-Institut für Sozialwissenschaften

Empfohlene Zitierung / Suggested Citation:
Schimpl-Neimanns, B., & Rendtel, U. (2001). *SAS-, SPSS- und STATA-Programme zur Berechnung der Varianz von Populationsschätzern im Mikrozensus ab 1996.* (ZUMA-Methodenbericht, 2001/04). Mannheim: Zentrum für Umfragen, Methoden und Analysen -ZUMA-. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-48759-3

ZUMA-Methodenbericht 2001/04

SAS-, SPSS- und STATA-Programme zur Berechnung der Varianz von Populationsschätzern im Mikrozensus ab 1996

Bernhard Schimpl-Neimanns& Ulrich Rendtel

Juni 2001

ISSN 1437-4129

ZUMA
Quadrat B2,1
Postfach 12 21 55
68072 Mannheim
Telefon: 0621-1246-263
Telefax: 0621-1246-100
E-mail: Schimpl-Neimanns@zuma-mannheim.de

Prof. Dr. Ulrich Rendtel
J.W. Goethe-Universität Frankfurt am Main
Fachbereich 02 – Institut für Statistik und Mathematik
Mertonstr. 17
60054 Frankfurt
Telefon: 069-798-237 55
Telefax: 069-798-236 35
E-mail: Rendtel@em.uni-frankfurt.de

Zusammenfassung

Erstmals enthält das Scientific Use File des Mikrozensus 1996 Stichprobeninformationen, die eine Berechnung der Varianz von Populationsschätzern ermöglichen. In diesem Bericht wird dokumentiert, wie mit den Standard-Software Programmen SAS, SPSS und STATA die Schätzung von Totals, Anteils- und Mittelwerten umgesetzt werden kann. Darüber hinaus werden Programme für die Hochrechnung von Ergebnissen nach der Anpassung an die Bevölkerungsfortschreibung vorgestellt.

1 Einleitung*

Bei der Schätzung von Populationswerten und ihren Varianzen ist die Ziehung der Stichprobe bzw. das Stichprobendesign zu berücksichtigen. Die in den Statistikpaketen verfügbaren Standard-Verfahren gehen bei der Varianzschätzung von der Annahme einer uneingeschränkten Zufallsauswahl aus. Das Scientific Use File des Mikrozensus, das der Forschung als faktisch anonymisierte 70%-Substichprobe ab der Erhebung 1989 zur Verfügung steht, ist jedoch eine geschichtete Klumpenstichprobe. Eine Klumpenstichprobe ist dadurch gekennzeichnet, dass die Auswahleinheit – im Mikrozensus ein Zählbezirk – alle Erhebungseinheiten umfasst. Die Annahme, dass die Beobachtungen unabhängig voneinander ausgewählt worden sind, trifft bei Klumpenstichproben somit nicht zu. In der Regel besitzen die Merkmale bei Klumpenstichproben im Vergleich zu uneingeschränkten Zufallsauswahlen eine größere Varianz. Bei Varianzschätzungen, die unter der Annahme einer einfachen Zufallsstichprobe durchgeführt werden, wird deshalb der Stichprobenfehler unterschätzt.

Informationen über die Schichtung und Klumpung stehen für die Forschung erstmals im Mikrozensus 1996 zur Verfügung. Wie Varianzschätzungen unter Berücksichtigung des Stichprobendesigns mit dem Scientific Use File des Mikrozensus ab 1996 durchgeführt werden können und welche Unterschiede zu den Fehlerrechnungen der statistischen Ämter bestehen, ist bereits an anderer Stelle diskutiert worden (Rendtel/Schimpl-Neimanns 2000, 2001). Dieser Methodenbericht baut darauf auf und konzentriert sich auf die praktische Umsetzung mit den Statistikpaketen SAS, SPSS und STATA. Die Programme werden im Folgenden dokumentiert und ihre Anwendung mit Beispielen beschrieben. Im Einzelnen werden behandelt: Die Schätzung von Gesamtwerten (Totals) sowie von Anteils- und Mittelwerten. Für die Schätzung von hochgerechneten Gesamtwerten, die nicht mit Standardprozeduren durchführbar

* Diese Arbeit geht zurück auf einen Gastaufenthalt des Zweitautors im Oktober 1999 bei ZUMA.

sind, wird ergänzend gezeigt, wie die Anpassung der Mikrozensus-Fallzahlen an die Bevölkerungsfortschreibung mit Hilfe einer Regressionsschätzung umgesetzt werden kann.

2 Eine Kurzdarstellung des Erhebungsdesigns des Mikrozensus und der Ziehung der 70%-Substichprobe

Als Ausgangspunkt der programmtechnischen Umsetzung von Varianzschätzungen ist zunächst die Stichprobe des Mikrozensus ab 1990 und das der Forschung zur Verfügung stehende Scientific Use File des Mikrozensus 1996 zu beschreiben, für das im Folgenden das Kürzel FAMZ für faktisch anonymisierter Mikrozensus verwendet wird. Die wichtigsten Eigenschaften werden in Übersicht 1 kurz dargestellt (vgl. Heidenreich 1994; Meyer 1994; Statistisches Bundesamt 1999). Im Anschluss daran wird die verwendete Notation beschrieben.

Übersicht 1: Zum Erhebungsdesign des faktisch anonymisierten Mikrozensus 1996

Stichprobeneigenschaften	**Mikrozensus ab 1990 – Originalmaterial**
Erhebungseinheiten	Haushalte, Personen
Auswahlgrundlage	Alte Bundesländer: Volkszählung 1987; Neue Bundesländer/Ost-Berlin (ab 1991): Bevölkerungsregister Statistik 1991 Aktualisierung der Stichprobe unter Berücksichtigung der Neubautätigkeit
Auswahlverfahren	Einstufig geschichtete Klumpenstichprobe
• **Schichtung**	Bundesland, Regierungsbezirk, Anpassungsschicht, Regionalschicht, Gebäudeschicht
• **Auswahleinheiten**	Primäreinheiten (PSU's): Auswahlbezirke PSU's sind Klumpen von i.d.R. zusammenliegenden Gebäuden bzw. Gebäudeteilen. Ein Auswahlbezirk verbleibt vier Jahre in der Stichprobe. In jedem Jahr scheidet ¼ der Auswahlbezirke aus (rotierendes Panel). Bildung der PSU's der Grundauswahl nach der Gebäudegröße (Gebäudeschicht): 1-4, 5-10, 11+ Wohnungen, Gemeinschaftsunterkünfte. Modifikationen der Gebäudeschicht bei Neubauauswahl: 1-4, 5-8, 9+ Wohnungen.
• **Auswahltechnik**	Grundauswahl: 1. Sortierung der PSU's nach regionaler Schichtuntergruppe, Kreis, Gemeindegrößenklasse und PSU-Nr. 2. Zusammenfassung von jeweils 100 aufeinander folgenden PSU's zu einer Zone. 3. Zufällige Zuordnung der PSU's einer Zone zu den Zahlen 0-99 (=“Stichprobennummer“). Anschließend Zusammenfassung der PSU's mit gleicher Stichprobennummer in 100 1%-Stichproben. 4. Zufällige Zuordnung von je 4 aufeinander folgenden Zonen (=“Block“) zu den Zahlen 1-4 zur Zerlegung der 1%-Stichproben in 4 Rotationsviertel á 0,25%. 5. Ermittlung von 20 1%-Stichproben durch zufällige Ziehung der Ordnungsnummern der Stichproben. Zufällige Ziehung der ersten Stichprobe für 1990. Die Grundauswahl (1-5) kann zusammenfassend als uneingeschränkte Zufallsauswahl beschrieben werden. Aktualisierung/Neubauauswahl: 6. Sortierung nach Aktualisierungsjahr und regionaler Kennung. Systematische Auswahl mit Zufallsstart.

Stichprobeneigenschaften	**Mikrozensus ab 1990 – Originalmaterial**
Stichprobenumfang	ca. 370.000 Haushalte; ca. 819.000 Personen (hochgerechnete, an die Bevölkerungsfortschreibung angepasste Fallzahlen; s.u.)
Auswahlsatz	1 Prozent
Hochrechnung	Zweistufiges Verfahren: 1. Kompensation der bekannten Ausfälle auf Haushaltsebene in 401 regionalen Untergruppen für jeweils 19 Merkmalskombinationen. 2. Anpassung der Stichprobenergebnisse an Eckzahlen aus der laufenden Bevölkerungsfortschreibung auf der Ebene regionaler Anpassungsschichten. Die Anpassungsklassen werden dabei gebildet durch die Angaben über die Zahl von Deutschen und Ausländern in der Gliederung nach Geschlecht. Die Anpassung für Soldaten und Wehrpflichtige erfolgt getrennt auf Regierungsbezirks- bzw. Landesebene auf Basis von Bestandsmeldungen des Verteidigungs- bzw. Innenministeriums. Die endgültigen Hochrechnungsfaktoren ergeben sich aus der Multiplikation des haushaltsbezogenen Kompensations- und des personenbezogenen Anpassungsfaktors.

Stichprobeneigenschaften	**Scientific Use File des Mikrozensus 1996 (faktisch anonymisierte 70%-Substichprobe) – Unterschiede zum Originalmaterial**
Auswahlverfahren	Zweiphasiges Ziehungsverfahren: 1. Phase: Ziehung der Haushalte wie in MZ-Originalauswahl (s.o.) 2. Phase: Ziehung der 70%-Substichprobe von Haushalten wie folgt
• **Schichtung durch Anordnung**	a) Sortieren der Datensätze nach Bundesland, Regierungsbezirk, Gemeindegrößenklasse, Zahl der Personen in Privathaushalten, Auswahlbezirks-Nummer, Nummer des Haushalts im Auswahlbezirk. (Datensätze für leer stehende Wohnungen und ausgefallene Haushalte werden vor der Sortierung gelöscht.)
• **Auswahltechnik**	Systematische Zufallsauswahl: b) Nach der Sortierung werden Haushalte fortlaufend nummeriert. Hierbei werden Anstaltspersonen wie Einpersonenhaushalte behandelt. (c) Ziehen/Löschen aller Sätze, deren letzte Platzziffer der Haushaltsnummer nicht einer von sieben ganzzahligen Zufallszahlen (z = 0, 1, 3, 4, 6, 7, 8) entspricht. D.h. Übernahme von 70% der Haushalte in die Substichprobe. d) Die Auswahlbezirks- und Haushaltsnummern werden nach der Substichprobenziehung aufs Neue fortlaufend nummeriert (EF3, EF4).
Stichprobenumfang	229.221 Haushalte; 509.243 Personen (nicht hochgerechnete Fallzahlen)
Hochrechnung	Das Scientific Use File wurde nicht extra an die laufende Bevölkerungsfortschreibung angepasst. Es enthält die MZ-Hochrechnungsfaktoren für Personen und Haushalte/Familien, die Ergebnisse des oben beschriebenen zweistufigen Verfahrens abbilden (EF750, EF751, EF755).

Im Scientific Use File liegen nicht alle Informationen für die Varianzberechnung vor. Von den Schichtungsmerkmalen der ersten Auswahlstufe können nur die Merkmale Bundesland, Gemeindegrößenklasse und Gebäudeschicht verwendet werden. Die Klumpenidentifikation ist mit der Variablen EF3 Auswahlbezirksnummer gegeben. Da im vorliegenden Datensatz keine Information über die Zahl der Haushalte in einem PSU vor der Ziehung der Substichprobe vorliegt, muss von einem fixen Auswahlsatz von 70 Prozent ausgegangen werden.

Es wird vereinfachend ein 2-stufiges Auswahlverfahren mit einer einfachen Auswahl auf jeder Stufe angenommen, wobei auf der 1. Stufe die Auswahl der PSU's und auf der 2. Stufe die Auswahl von jeweils 70 Prozent der Haushalte einer PSU erfolgt. In der im Folgenden verwendeten Notation von Särndal et al. (1992): „simple random sampling without replacement" (SI, SI).

Die Laufindizes lauten:

Schichten: $h \in \{1,...,H\}$

PSU's: $i \in \{1,...,N_h\}$

Haushalte: $k \in \{1,...,N_i\}$

Der Merkmalswert $y_{h,i,k}$ bezieht sich somit auf den Haushalt k in PSU i in Schicht h. Die Anzahl der Primäreinheiten der Grundgesamtheit in der Schicht h ist $N_{I,h}$. Die Stichprobe der PSU's in den H Schichten hat den Umfang $n_{I,h}$, die Stichprobe der Haushalte aus PSU i den Umfang n_i. Weiterhin bezeichnet $\pi_{I,i}$ die Ziehungswahrscheinlichkeit von PSU i und $\pi_{k|i}$ die bedingte Ziehungswahrscheinlichkeit von Haushalt k aus PSU i, wenn PSU i gezogen wurde. Im Falle des Mikrozensus gilt $\pi_{I,i} = 0{,}01$ und für den FAMZ wird ein fester Auswahlsatz von $\pi_{k|i} = 0{,}7$ angenommen.

3 Die Schätzung von Totals

Bei der Schätzung des Gesamtaufkommens (Total) t eines Merkmals y wird der π-Schätzer verwendet, der auf dem Kehrwert der Ziehungswahrscheinlichkeiten basiert. Man enthält als Schätzung für das Gesamttotal:

$$\hat{t} = \sum_{h=1}^{H} \hat{t}_h = \frac{100}{0{,}7} \sum_{k \in s} y_k$$

Bei einem 2-stufigen Ziehungsverfahren ergibt sich die Varianz unter Verwendung der Stichprobenwerte durch (vgl. Särndal et al. 1992: 142):

$$V_{SI,SI}\left(\hat{t}_h\right) = N_{I,h}^2 \frac{1-f_h}{n_{I,h}} S_{s_{I,h}}^2 + \frac{N_{I,h}}{n_{I,h}} \sum_{i \in U_{I,h}} N_i^2 \frac{1-f_i}{n_i} S_{s_i}^2$$

wobei $f_h = \frac{n_{I,h}}{N_{I,h}}$ und $f_i = \frac{n_i}{N_i}$

$$S^2_{s_{I,h}} = \frac{1}{n_{I,h}-1}\sum_{i\in s_{I,h}}\left(\hat{t}_i - \hat{\bar{t}}_{s_{I,h}}\right)^2 \quad = \text{"Varianz between PSU"}$$

$$S^2_{s_i} = \frac{1}{n_i-1}\sum_{k\in s_i}\left(y_k - \bar{y}_{s_i}\right)^2 \quad = \text{"Varianz within PSU"}$$

Hierbei ist $\hat{t}_i$ das geschätzte Total für PSU i und $\hat{\bar{t}}_{s_{I,h}}$ der Stichprobenmittelwert der $\hat{t}_i$ in der Schicht h. $\bar{y}_{s_i}$ ist der Stichprobenmittelwert der y_k in PSU i.

Die erwartungstreue Schätzung der Varianz $\hat{V}_{SI,SI}$ vereinfacht sich auf:

$$\hat{V}_{SI,SI} = 100^2 \times 0{,}99 \times n_{I,h} \times \text{"Varianz between PSU"} + 100\sum_{i\in U_{I,h}}\frac{0{,}3}{0{,}7^2} n_i \times \text{"Varianz within PSU"}$$

In der Praxis können die notwendigen Berechnungen mit allen gängigen Statistikpaketen durchgeführt werden. Für die Summenbildung auf den verschiedenen Ebenen Haushalt, PSU und Schicht ist lediglich die Datenaggregation und die Berechnung von Varianzen bzw. Standardabweichungen erforderlich.

Die im Anhang dokumentierten Programme enthalten bereits die Konstruktion einer aus den Merkmalen Bundesland (Variable ef1), Gemeindegrößenklasse (ef708) und Gebäudeschicht (ef712) kombinierten Schichtvariablen (schicht) und alle nötigen Berechnungsschritte für die nachfolgende Beispielauswertung. Vom Anwender ist lediglich das interessierende Merkmal (y) und die Subpopulation (z) zu definieren. Änderungen, die bei anderen als für dieses Beispiel verwendeten Merkmalsabgrenzungen notwendig werden, sind mit spitzen Klammern <> gekennzeichnet. Bei haushaltsbezogenen Merkmalen muss ggf. vorab auf Haushaltsebene aggregiert werden. Unter Umständen kann dann der personenbezogene Programmteil A entfallen oder gekürzt werden. Wie die Berechnungen im Einzelnen mit SAS, SPSS und STATA umgesetzt werden können, ist den Dokumentationen zu den VarMZ_T.* Programmen im Anhang zu entnehmen. Die verschiedenen Programmschritte sind jeweils kurz kommentiert. In diesen Programmen werden Totals, Varianzen bzw. Standardfehler sowie der relative Standardfehler (=Variationskoeffizient) und der Design-Effekt Faktor[1] berechnet.

Am Beispiel des Merkmals „Ledige“ werden für die Subpopulation „Bevölkerung in Privathaushalten“ folgende Ergebnisse ermittelt:

[1] Der Design-Effekt Faktor beschreibt das Verhältnis des Design basierten Standardfehlers im FAMZ zum Standardfehler einer Stichprobenziehung gleichen Umfangs unter der Annahme einer uneingeschränkten Zu-

Tabelle 1: Schätzergebnisse für das Merkmal Ledige Bevölkerung in Privathaushalten

Kennwert	Programme VarMZ_T.* (s. Anhang)	SAS Prozedur surveymeans	STATA Prozedur svytotal
Total (in 1000)	27.323,6	27.323,6	27.323,6
Standardfehler (in 1000)	85,4	85,3	85,3
Between-Komponente	85,3		
Within-Komponente	3,4		
relativer Standardfehler (in %)	0,31		
Design-Effekt Faktor *)	1,74		1,73

*) Bei Verwendung der STATA Prozedur svytotal: √(Deff)

Die Berechnung der Within Varianzkomponente von $\hat{V}_{SI,SI}$ erfordert die Ermittlung der Standardabweichung der y-Werte innerhalb von ca. 40.000 PSU's. Bei einer sehr kleinen Auswahlwahrscheinlichkeit der Primäreinheiten (1. Stufe) reicht die Berechnung der Between-Varianz als Näherung aus, so dass auch die Standardprozeduren von SAS und STATA verwendet werden können, in denen nur die Between-Varianz berechnet wird. Der Vergleich der geschätzten Standardfehler des Beispielmerkmals in Tabelle 1 zeigt, dass die Within-Varianz im Vergleich zur Between-Varianz sehr klein ist und vernachlässigt werden kann. Die in Tabelle 1 berichteten vereinfachten Varianzberechnungen können mit SAS und STATA wie folgt durchgeführt werden.

SAS-Nutzer können ab der Version 7 für die Design basierte Varianzschätzung auf die Prozedur surveymeans zurückgreifen. Eine korrekte Schätzung für Subpopulationen (Domains) ist aber erst ab Version 8.1 implementiert. Für Nutzer der Version 8.0 stellt SAS ersatzweise das Macro SMSUB zur Verfügung.[2] Die folgende Verwendung der Prozedur surveymeans geht davon aus, dass im SAS-Datensatz die aus Bundesland, Gebäudeschicht und Gemeindegrößenklasse kombinierte Schichtvariable „schicht", die Ordnungsnummern der Auswahlbezirke (PSU) „ef3", das y-Merkmal (ledige Bevölkerung in Privathaushalten) und die Hochrech-

fallsauswahl (vgl. Rendtel/Schimpl-Neimanns 2001).

[2] Siehe SAS Institute: FAQ. Can PROC SURVEYMEANS compute subgroup (domain) variance estimates? <http://www.sas.com/service/techsup/faq/stat_proc/surveymeansproc1689.html>; letzter Zugriff: 29.01.2001. Zur Verwendung des SMSUB-Macros muss im SAS-Datensatz eine Variable vorliegen, welche die Anzahl der PSU's pro Schicht in der Grundgesamtheit enthält (Option POPSIZE). Bei verschiedenen Testauswertungen konnte das Macro aus ungeklärten Gründen die Varianz nicht berechnen. Wir verzichten deshalb auf eine

nungskonstante „gew“ (=100/0,7) bereits vorliegen. Der Auswahlsatz des Mikrozensus ist als Option (RATE=0.01) anzugeben. Die entsprechenden Angaben, mit denen Schätzungen für Totals zum obigen Beispiel durchgeführt werden, lauten dann:[3]

```
PROC SURVEYMEANS DATA=<library.dateiname> RATE=0.01 SUM ;
    STRATA schicht;
    VAR y;
    CLASS y;
    CLUSTER ef3;
    WEIGHT gew;
    ODS OUTPUT STATISTICS = <ergebnis>;
RUN;
```

Die Varianzschätzung auch mit STATA sehr einfach durchzuführen.[4] Gegeben, dass die entsprechenden Variablen „schicht“ etc. vorliegen, lauten die Programmanweisungen für die Prozedur svytotal:

```
SVYSET STRATA schicht
SVYSET PSU ef3
GENERATE gew=100/0.7
SVYSET PWEIGHT gew
GENERATE r=0.01
SVYSET FPC r
SVYTOTAL y, deff
```

4 Schätzung von Verhältnis- und Mittelwerten

Bei der Schätzung des Verhältnisses $\hat{R} = \hat{t}_y / \hat{t}_z$ der Fallzahlen von zwei Merkmalen y und z wird zur Berechnung der asymptotischen Varianz von $\hat{R}$ eine Taylorentwicklung der Funktion $f(t_y, t_z) = t_y / t_z$ benutzt:

$$f(\hat{t}_y, \hat{t}_z) \approx f(t_y, t_z) + \frac{\partial f}{\partial t_y}\Big|_{t_y}(\hat{t}_y - t_y) + \frac{\partial f}{\partial t_z}\Big|_{t_z}(\hat{t}_z - t_z) = \text{Konstante} + \frac{1}{t_z}\sum_{k \in s}\frac{u_k}{\pi_k}$$

wobei $u_k = (y_k - R\, z_k)$.

Dokumentation zur Verwendung des Macros SMSUB.

[3] Die hier verwendete Option CLASS setzt ein qualitatives y-Merkmal voraus.

[4] Bei der Gebäudeschicht „Gemeinschafts-/Anstaltsunterkunft“ (EF712=4) ist im FAMZ häufig nur eine PSU pro Schicht vorhanden, so dass für diese Schichten eine Varianzschätzung nicht ohne Weiteres durchführbar ist. Gegebenenfalls sind diese PSU’s vorab aus der Berechnung auszuschließen oder mit anderen PSU’s zusammenzufassen. Eine Liste von solchen Problemfällen kann mit dem STATA-Kommando svydes erstellt werden.

Zur Berechnung von $V(\hat{t}_u)$ können die für die Schätzung der Varianz von Totals benutzten Formeln und Programme verwendet werden. Es ist lediglich y_k durch die Hilfsgröße $u_k = y_k - \hat{R}z_k$ zu ersetzen und am Schluß der Berechnungen der Varianzen ist durch $\hat{t}_z^2$ zu dividieren.

Bei Verwendung der Programme VarMZ_R.* sind nur je nach Anwendungszweck das Zähler- und Nennermerkmal zu definieren. Alle weiteren Berechnungsschritte sind in den Programmen enthalten und analog dem ersten Programm zur Schätzung von Totals VarMZ_T.* umgesetzt (s.o.). Am Beispiel des bereits verwendeten y-Merkmals „Ledig“ und des z-Merkmals „Bevölkerung in Privathaushalten“ bzw. des Anteils lediger Personen an der Bevölkerung in Privathaushalten werden folgende Ergebnisse ermittelt:

Tabelle 2: Schätzergebnisse für den Anteil Lediger an der Bevölkerung in Privathaushalten

Kennwert	Programme VarMZ_R.* (s. Anhang)	SAS Prozedur surveymeans	STATA Prozedur svymean
Anteil R (in Prozent)	37,96	37,96	37,96
Standardfehler (in Prozent)	0,0726	0,0725	0,0726
Between-Komponente	0,0726		
Within-Komponente	0,0034		
relativer Standardfehler (in %)	0,1914		
Design-Effekt Faktor *)	1,06		1,06

*) Bei Verwendung der STATA-Prozedur svymean: √(Deff)

Bei Verwendung der Prozedur surveymeans von SAS ist bei Anteilen oder Mittelwerten statt des Statistics-Schlüsselworts SUM das Schlüsselwort MEANS zu verwenden. In STATA stehen für die Berechnung der Varianz von Anteilen neben svymeans mit svyprop und svyratio weitere Prozeduren zur Verfügung.

Der **Populationsmittelwert** zu einem Merkmal y kann als das gewichtete Stichprobenmittel $\hat{\bar{y}}_s$ berechnet werden (vgl. Särndal et al. 1992: 182)

$$\hat{\bar{y}}_s = \frac{\sum_{k \in s} Y_k / \pi_k}{\sum_{k \in s} 1/\pi_k}$$

In dieser Form ist $\hat{\bar{y}}_s$ ein Spezialfall von $\hat{R}$ mit einem dichotomen z-Merkmal $z_k = \{0,1\}$, so dass die Programme VarMZ_R.* zur Varianzschätzung für Anteilswerte ohne größere Änderungen benutzt werden können. Die Programmänderungen sind im Anhang für die Beispielauswertung zum mittleren Heiratsalter von Frauen dokumentiert.

Tabelle 3: Schätzergebnisse für das mittlere Heiratsalter von Frauen

Kennwert	Programme VarMZ_R.* (s. Anhang)	SAS Prozedur surveymeans	STATA Prozedur svymean
Mittelwert	24,73	24,73	24,74
Standardfehler	0,0202	0,0201	0,0201
Between-Komponente	0,0201		
Within-Komponente	0,0011		
relativer Standardfehler (in %)	0,08		

Variablendefinition: siehe Anhang

5 Die Varianz von Populationsschätzern nach der Anpassung an die Bevölkerungsfortschreibung

Die Nutzer der Scientific Use Files können für die so genannte gebundene Hochrechnung auf die im Datensatz enthaltenen Hochrechnungsfaktoren für Personen und Haushalte zurückgreifen (s. Übersicht 1). Die Verwendung von Gewichten, die im Wesentlichen aus der Anpassung der MZ-Fallzahlen an die Bevölkerungsfortschreibung resultieren, kann als Regressionsschätzung interpretiert werden. Der hier benutzte Regressionsschätzer $\hat{t}_{reg}$ basiert auf dem Group Mean Modell (vgl. Rendtel/Schimpl-Neimanns 2001):

$$\hat{t}_{reg} = \sum_{k \in U} \hat{y}_k = \sum_{g=1}^{G} \sum_{k \in U_g} \hat{B}_g = \sum_{g=1}^{G} N_g \hat{B}_g = \sum_{g=1}^{G} \sum_{k \in s_g} \frac{N_g}{\hat{N}_g} \cdot \frac{y_k}{\pi_k} = \sum_{g=1}^{G} \sum_{k \in s_g} w_g \frac{y_k}{\pi_k}$$

Der Faktor $w_g = N/\hat{N}$ beschreibt das Verhältnis von N_g = Umfang von Gruppe g in der Grundgesamtheit (= Soll - Vorgabe) zu $\hat{N}_g$ = geschätzter Umfang von Gruppe g (= Ist - Wert). Der Regressionsschätzer lässt sich damit als ein "gewichtetes Mittel" darstellen:

$$\hat{t}_{reg} = \sum_{k \in s} w_k \frac{y_k}{\pi_k}$$

Wird für die Herleitung von $V(\hat{t}_{reg})$ wieder eine Taylorentwicklung von $\hat{t}$ benutzt, erhält man eine asymptotische Näherung für $V(\hat{t}_{reg})$. Der lineare Teil der Taylorentwicklung ist durch die folgende Hilfsgröße u_k gegeben (vgl. Särndal et al. 1992: 331):

$$u_k = \frac{N_g}{\hat{N}_g}(y_k - \hat{B}_g) = w_k(y_k - \bar{y}_{s_g}) \qquad k \in s_g$$

Dieser lineare Anteil ist also die mit w_k gewichtete Abweichung des Merkmalswerts y_k von dem jeweiligen Gruppenmittelwert $\bar{y}_{s_k}$ in der Stichprobe. Als asymptotische Varianz wird die Varianz dieses Hilfsmerkmals u verwendet. Bei der praktischen Berechnung hat man wieder lediglich y_k durch u_k in den Gleichungen zur Schätzung der Totals zu ersetzten.

Die Beispielprogramme VarMZ_A.* greifen auf den Personen-Hochrechnungsfaktor (EF750) zurück. Die Programme VarMZ_A.* sind im Vergleich zur Varianzschätzung von Totals und Ratios rechentechnisch etwas aufwendiger, da die Daten bei der Berechnung der Regressionskoeffizienten zusätzlich auf der Gruppenebene (gruppe) aggregiert werden müssen. Die Gruppen sind als Kombination der Variablen Bundesland (ef1) und Anpassungsklasse (anp) definiert. Da im Scientific Use File die Anpassungsschichten nicht identifizierbar sind, steht als Regionalinformation für die Gruppenbildung lediglich das Bundesland zur Verfügung. In den Programmen VarMZ_A.* ist die Umsetzung der Anpassungsklasse auf Personenebene sowie die Schichtkonstruktion bereits enthalten. Wie in den anderen Programmen sind vom Anwender noch das interessierende Merkmal und die Subpopulation zu definieren. Zusätzlich ist zu beachten, dass bei Auswertungen auf Haushalts- oder Familienebene, die in der Regel unter Verwendung des Haushalts-Hochrechnungsfaktors (Variable EF751) durchgeführt werden,[5] entsprechende Änderungen bei der Definition der Variablen Soll-durch-Ist (soll_ist) vorgenommen werden müssen. Für die Abgrenzung der Anpassungsklasse (anp) können in

[5] Wir weichen beim Beispiel in Tabelle 4 von der Praxis der statistischen Ämter ab, die für Auswertungen der Bevölkerung in Privathaushalten den Haushalts-Hochrechnungsfaktor verwenden, und benutzen statt dessen den Personen-Hochrechnungsfaktor.

diesem Fall die Eigenschaften der Bezugsperson des Haushalts (EF507=1) herangezogen werden.[6]

Tabelle 4: Schätzergebnisse für das Merkmal Ledige Bevölkerung in Privathaushalten mit Anpassung an die Bevölkerungsfortschreibung

Kennwert	Programme VarMZ_A.* (s. Anhang)	STATA Prozedur svymean*
Total (in 1000)	31.360,1	31.360,1
Standardfehler (in 1000)	60,8	60,7
Between-Komponente	60,7	
Within-Komponente	2,8	
relativer Standardfehler (in %)	0,19	

*) Berechnung des Standardfehlers unter Verwendung des Hilfsmerkmals u

Die Ergebnisse in Tabelle 4 für das eingangs verwendete Beispielmerkmal zeigen mit dem relativen Standardfehler von 0,19 Prozent eine deutliche Reduktion der Varianz gegenüber dem Wert von 0,31 Prozent bei der Schätzung ohne Anpassung (siehe Tabelle 1). In den meisten der von uns durchgeführten Schätzungen fallen die Varianzreduktionen zumeist aber wesentlich geringer aus.

In den Standardprozeduren der hier verwendeten Statistikpakete ist diese Regressionsschätzung für die Behandlung der an die Bevölkerungsfortschreibung angepassten Mikrozensus-Fallzahlen nicht implementiert. Ersatzweise können aber Nutzer von SAS und STATA die Prozeduren für die Schätzung von Totals heranziehen und das Hilfsmerkmal u für die Varianzschätzung einsetzen. Das angepasste Total muss jedoch gesondert berechnet werden. Diese Lösung wird im Folgenden für STATA dokumentiert, wobei vorausgesetzt wird, dass die Berechnungsschritte des im Anhang dokumentierten Programms VarMZ_A.DO bereits bis einschließlich der Berechnung des Hilfsmerkmals u (generate u=soll_ist*(y-B_dach)) durchgeführt worden sind. Man erhält damit das in Tabelle 4 berichtete Ergebnis.

```
(...)
generate u=soll_ist*(y-B_dach)
/* soll_ist = ef750 Personen-Hochrechnungsfaktor */
/* Berechnung und Ausgabe des gewichteten Totals (in 1000) */
egen y_g=sum(soll_ist*y/7)
list y_g in 1/1
```

[6] Die betreffenden Variablen sind: Geschlecht (EF577), Staatsangehörigkeit (EF559), Soldat/Wehrpflichtiger (EF564=9, 10).

```
/* Sortieren nach Schicht-, PSU- und Haushaltsnummer */
sort schicht psu hhnr
svyset strata schicht
svyset psu psu
generate gew=100/0.7
svyset pweight gew
generate r=0.01
svyset fpc r
svytotal u
```

6 Schluß

Mit Hilfe der in diesem Bericht beschriebenen Programme und Vorgehensweisen bei Verwendung von Standardprozeduren der Statistikprogramme SAS und STATA ist eine statistisch angemessene Varianzschätzung ab dem Scientific Use Files des Mikrozensus 1996 möglich. Damit ist es insbesondere nicht mehr nötig, den Stichprobenfehler unter der Annahme einer uneingeschränkten Zufallsauswahl zu berechnen und mit Hilfe der vom Statistischen Bundesamt veröffentlichten „Zuschlagsfaktoren“ für Design-Effekte zu korrigieren. Die Verwendung der für den Mikrozensus berichteten Design-Effekte führt zu einer Überschätzung des Standardfehlers im Vergleich zur direkten Varianzschätzung (Rendtel/Schimpl-Neimanns 2001). Die direkte Varianzschätzung ist deshalb für das Scientific Use File immer zu präferieren.

Berücksichtigt man, dass über 500.000 Fälle zu verarbeiten sowie zu sortieren sind und die Daten in verschiedenen Rechenschritten auf Haushalts-, PSU- und Schichtebene aggregiert werden, stellt sich abschließend die Frage nach dem Aufwand und den Laufzeiten der Programme. Hierzu werden in Tabelle 5 grobe Näherungswerte berichtet, die bei der Auswertung für das Beispielmerkmal „Ledige Bevölkerung in Privathaushalten“ ermittelt wurden. Beim Vergleich der Laufzeiten ist zu erkennen, dass die Statistikpakete den vorhandenen Arbeitsspeicher unterschiedlich nutzen. Bei einem geeigneten Rechner liegt selbst die Laufzeit von 8 Minuten bei SAS nicht besonders hoch.[7] STATA schneidet unter den verwendeten Programmen auf Grund einer effizienteren Speicherwaltung mit ca. 2 Minuten am besten ab. Benutzer von STATA können bei der Varianzschätzung darüber hinaus auf weitere komfortable Prozeduren zurückgreifen. Einschränkungen der Nutzbarkeit der SAS Prozedur Surveymeans bestehen bei der Version 8.0 hinsichtlich der nicht korrekten Varianzschätzung für Subpopulationen.

[7] SAS benötigt alleine für das Einlesen des SAS-Datenfiles rund 6 Minuten.

Tabelle 5: Vergleichsübersicht der Laufzeiten von Programmen zur Varianzschätzung von Totals*

Statistiksoftware	Programm	Minuten: Sekunden (Real time; ca. Werte)
SAS Version 8.0 for Windows	VarMZ_T.SAS	8:00
	SAS surveymeans	7:00
SPSS for Windows Release 9.0.0	VarMZ_T.SPS	4:40
STATA 6.0 for Windows**	VarMZ_T.DO	1:40
	STATA svytotal	2:00

* Verwendeter PC: Pentium III MMX; 128 MB Arbeitsspeicher; 450 Mhz; Windows NT 4.0, SP6.
** Bei den Testläufen wurde STATA 80 MB Arbeitsspeicher zugewiesen.

Zusammenfassend kann festgehalten werden, dass Varianzschätzungen für das Scientific Use Files des Mikrozensus ab 1996 recht einfach und unter Verwendung von Standard-Statistiksoftware durchzuführen sind. Die im File enthaltenen Schicht- und Klumpenidentifikatoren ermöglichen es den Nutzern erstmals, eine wesentliche Qualität dieses Datensatzes, nämlich die auf Grund des hohen Stichprobenumfangs sehr niedrigen Stichprobenfehler der Schätzergebnisse, effizient auszuschöpfen und dem Design angemessene Schätzungen selbst durchzuführen.

Literatur

Heidenreich, Hans-Joachim, 1994: Hochrechnung des Mikrozensus ab 1990. S. 112-123 in: *Gabler, Siegfried, Jürgen H.P. Hoffmeyer-Zlotnik und Dagmar Krebs* (Hg.*)*: Gewichtung in der Umfragepraxis. Opladen: Westdeutscher Verlag.

Meyer, Kurt, 1994: Zum Auswahlplan des Mikrozensus ab 1990. S. 106-111 in: *Gabler, Siegfried, Jürgen H.P. Hoffmeyer-Zlotnik und Dagmar Krebs* (Hg.*)*: Gewichtung in der Umfragepraxis. Opladen: Westdeutscher Verlag.

Rendtel, Ulrich, und Bernhard Schimpl-Neimanns, 2000: Varianzschätzungen für den faktisch anonymisierten Mikrozensus. Jahrbücher für Nationalökonomie und Statistik 220(6): 759-776.

Rendtel, Ulrich, und Bernhard Schimpl-Neimanns, 2001: Die Berechnung der Varianz von Populationsschätzern im Scientific Use File des Mikrozensus ab 1996. ZUMA-Nachrichten 48: 85-116. (URL http://www.gesis.org/Publikationen/Zeitschriften/-ZUMA_Nachrichten/documents/pdfs/zn48_10-bernhard.pdf)

Särndal, Carl-Erik, Bengt Swensson und Jan Wretman, 1992: Model Assisted Survey Sampling. New York: Springer.

Statistisches Bundesamt, 1999: Zum Auswahlplan des Mikrozensus ab 1990. Seite E2 49-56 in: Arbeitsunterlagen zum Mikrozensus. Das Erhebungsprogramm des Mikrozensus seit 1957. Wiesbaden (Loseblattsammlung; September 1999).

Anhang: SAS-, SPSS- und STATA-Programme

```
/* ------------------------- VarMZ_T.SAS --------------------------------
   1. Programmname: VarMZ_T.SAS (URL www.gesis.org/Dauerbeobachtung/
         Mikrodaten/mikrodaten_tools/Varianz/VarMZ_T.SAS )
   2. Programmautoren: Ulrich Rendtel (rendtel@em.uni-frankfurt.de)
         Bernhard Schimpl-Neimanns (schimpl-neimanns@zuma-mannheim.de)
   3. Zweck des Programms: Berechnung der Varianz des Pi-Schaetzers fuer
         ein Merkmal Y im faktisch anonymisierten Mikrozensus (FAMZ) 1996.
         Hier am Beispiel des Merkmals Y "Ledige" (EF35=1) fuer die
         Subpopulation Z "Bevoelkerung in Privathaushalten" (EF506=1)
   4. Weiterfuehrende Aufgabenbeschreibungen:
         Rendtel, Ulrich; Schimpl-Neimanns, Bernhard: Varianzschaetzungen
            fuer den faktisch anonymisierten Mikrozensus. In: Jahrbuecher
            fuer Nationaloekonomie und Statistik, 220/6, 2000, S. 759-776.
         Rendtel, Ulrich; Schimpl-Neimanns, Bernhard: Die Berechnung der
            Varianz von Populationsschaetzern im Scientific Use File des
            Mikrozensus. In: ZUMA-Nachrichten Nr. 48, 2001, S. 85-116
            (URL www.gesis.org/Publikationen/Zeitschriften/
            ZUMA_Nachrichten/documents/pdfs/zn48_10-bernhard.pdf )
         Schimpl-Neimanns, Bernhard; Rendtel, Ulrich: SAS-, SPSS- und
            STATA-Programme zur Berechnung der Varianz von
            Populationsschaetzern im Mikrozensus. ZUMA-Methodenbericht
            Nr. 2001/04. Mannheim. (URL www.gesis.org/Publikationen/
            Berichte/ZUMA_Methodenberichte/documents/pdfs/tb01_04.pdf )
   5. Projektbeginn: Oktober 1999
   6. Letzte Programmaenderung: 26. Januar 2001
   7. Programmstatus: Getestet mit SAS fuer Windows V8.0 (Windows NT 4.0,
         SP6) und Mikrozensus 1996 (faktisch anonymisierte 70%-
         Substichprobe).
   8. Erforderliche Programmeingaben: SAS-Datensatz basierend auf den
         Rohdaten des Mikrozensus. Das File sollte keine Missing Values
         enthalten.
         Schichtvariablen: EF1 Bundesland, EF708 Gemeindegroessenklasse,
            EF712 Gebaeudegroessenklasse
         Klumpenidentifikation (PSU): EF3 Auswahlbezirk
         Haushaltsidentifikation (HHNR): EF4 Haushaltsnummer
         Die bei anderen als in diesem Beispiel verwendeten Y-Variablen
         und Subpopulationen (Z) sowie insgesamt zu aendernden
         Programmschritte sind mit spitzen Klammern <> gekennzeichnet.
   9. Grobe Programmstruktur:
       A Einzeldaten einlesen und benoetigte Variablen definieren
         Berechnen der Haushaltstotals
       B Haushaltsbezogene Daten weiterverarbeiten (ggf. einlesen)
       C Berechnen der PSU Totals und PSU Within Varianzen
         Berechnen der Schicht Totals, der Between Varianz und der
            Summe der gewichteten Within Varianzen
         Summation der Totals und Varianz Terme ueber die Schichten
         Berechnung der Standard Abweichungen
         Berechnung der Varianz unter Annahme der Binomialverteilung
         Berechnung der auszugebenden Kennwerte
         Ausgabe der Kennwerte
   ---------------------------------------------------------------------- */

   /* Teil A: Daten Personenbezogen  */

LIBNAME <libname> '<Verzeichnis>';

DATA MZ_pers ;
     set <libname.filename>
        (keep = ef1 ef3 ef4 <ef35> <ef506> ef708 ef712);
     PSU=ef3;
     HHNR=ef4;
     schicht = ef1*100 + ef708*10 + ef712;
     y  = <(EF35 eq 1)*(ef506 eq 1)> ;
```

```
     z  = <(ef506 eq 1)> ;
     LABEL  PSU="Auswahlbezirksnummer (EF3)"
            HHNR="lfd. Haushaltsnummer im Auswahlbezirk (EF4)"
            SCHICHT="Bu.land * Gem.groesse * Geb.schicht"
            Y="<Ledige in Privathaushalten (1, sonst 0)>"
            Z="<Bevoelkerung in Privathaushalten (1, sonst 0)>";
     run;

   /* Sortieren falls noetig, sonst ueberspringen */
Proc sort data=MZ_pers;
     by schicht psu hhnr;
     run;

   /* Berechnen der Haushaltstotals */
Proc Means data=MZ_pers noprint;
     var y z;
     by schicht psu hhnr;
     output out=MZ
            sum=y_k z_k;
     run;
   /* Ende Teil A */


   /* Teil B: Daten Haushaltsbezogen
                   Beim Einlesen haushaltsbezogener Daten muessen
                   Y und Z als y_k und z_k aggregiert vorliegen */

    /* Sortieren falls noetig, sonst ueberspringen */
Proc sort data=MZ;
     by schicht psu hhnr;
    /* Ende Teil B */


    /* Teil C: Ab hier weiter in beiden Faellen  */

    /* Berechnen der PSU Totals und PSU Within Varianzen */
Proc Means data=MZ noprint;
     var y_k;
     by schicht psu ;
     output out=psu_data
            sum=psu_y
            Var=psu_var
            N  =psu_n;
     run;

    /* Berechnen der Schicht Totals, der Between Varianz und der
       Summe der gewichteten Within Varianzen */

Data PSU_data;
     set Psu_data;
        /* Einige PSUs sind im FAMZ nur mit E I N E M  Haushalt
           repraesentiert: Missings rekodieren */
     if Nmiss(PSU_Var)=1 then PSU_var=0;
        /* Berechnung n_i*S^2(s_i)   */
     N_Var=PSU_N*PSU_Var;

Proc means data=PSU_data noprint;
     Var PSU_y N_var;
     by Schicht;
     output out=str_data
            sum=STR_y STR_with
            Var=Between
            N  =STR_n ;
```

```
     run;

  /* Strata Varianz = 100^2*0.99*n_(I,h)*S^2(n_I,h)          [= V_betw ]
                     +100*0.3/(0.7*0.7) * Summe n_i*S^2(s_i) [= V_with ] */
Data STR_data;
     set STR_data;
     Total =100*STR_y/0.7;
     V_betw=10000*0.99*STR_N*Between/(0.7*0.7) ;
     V_with= 100*0.3*STR_with/(0.7*0.7);
     v=v_Betw + V_with ;

   /* Summation der Totals und Varianz Terme ueber die Strata */

Proc Means data=STR_data  noprint;
     Var total V_betw V_with V;
     output out = ergebnis
            sum = TOtAL V_betw V_with V;
     run;

   /* Berechnung der Standard Abweichungen */
data ergebnis ;
     set ergebnis ;
     sig_V=sqrt(V);
     sig_B=sqrt(V_betw);
     sig_W=sqrt(V_with);
     rel=sig_V/Total;

   /* Berechnung der Binomial Varianz
      hier: Y-Merkmal und Anzahl der zur Subpopulation gehoerenden
      Personen liegen auf Haushaltsebene aggregiert vor (y_k, z_k) */

Proc means data=MZ noprint;
     Var y_k z_k ;
     output out=binomial
            sum=y_stichp n_stichp;

data binomial;
     set binomial;
     p_dach=y_stichp/n_stichp ;

Data ausgabe;
     Merge binomial ergebnis;
     Rel_Bin=sqrt(0.99*(1- p_dach)/(p_dach*(n_stichp - 1)) );
     deft=rel/rel_bin;
     total = total/1000;    /* Ausgabe in 1000 */
     sig_v = sig_v/1000;
     sig_b = sig_b/1000;
     sig_w = sig_w/1000;
     rel = rel*100;         /* Ausgabe in Prozent */
     rel_bin = rel_bin*100;
     p_dach=p_dach*100;
     LABEL total="Total (in 1000)"
           sig_v="Std.Fehler (in 1000)"
           sig_b="Std.Fehler Between-Teil (in 1000)"
           sig_w="Std.Fehler Within-Teil (in 1000)"
           rel = "relativer Std.Fehler (in %)"
           deft = "Design-Effekt Faktor des Std.Fehlers"
           rel_bin="Relativer Std.Fehler Binomialverteilung (in %)"
           p_dach="Anteil Y in Subpopulation Z (in %)";

Proc Print data=ausgabe;
     Var total sig_V sig_B sig_W  rel deft rel_bin p_dach ;
     run;
```

```
/* ------------------------- VarMZ_R.SAS --------------------------------
   1. Programmname: VarMZ_R.SAS (URL www.gesis.org/Dauerbeobachtung/
         Mikrodaten/mikrodaten_tools/Varianz/VarMZ_R.SAS )
   2. Programmautoren: Ulrich Rendtel (rendtel@em.uni-frankfurt.de)
         Bernhard Schimpl-Neimanns (schimpl-neimanns@zuma-mannheim.de)
   3. Zweck des Programms: Berechnung der Varianz des Pi-Schaetzers fuer
         das Verhaeltnis R zweier Totals, t_y und t_z, im faktisch
         anonymisierten Mikrozensus (FAMZ) 1996.
         Hier am Beispiel des Merkmals Y "Ledige" (EF35=1) fuer die
         Subpopulation Z "Bevoelkerung in Privathaushalten" (EF506=1)
         Das Programm kann auch zur Berechnung der Varianz des
         arithmetischen Mittelwerts der Variablen Y fuer die Subpopulation
         Z verwendet werden. (Ein Beispiel befindet sich am Ende des
         Programms.)
   4. Weiterfuehrende Aufgabenbeschreibungen:
         Rendtel, Ulrich; Schimpl-Neimanns, Bernhard: Varianzschaetzungen
            fuer den faktisch anonymisierten Mikrozensus. In: Jahrbuecher
            fuer Nationaloekonomie und Statistik, 220/6, 2000, S. 759-776.
         Rendtel, Ulrich; Schimpl-Neimanns, Bernhard: Die Berechnung der
            Varianz von Populationsschaetzern im Scientific Use File des
            Mikrozensus. In: ZUMA-Nachrichten Nr. 48, 2001, S. 85-116
            (URL www.gesis.org/Publikationen/Zeitschriften/
            ZUMA_Nachrichten/documents/pdfs/zn48_10-bernhard.pdf )
         Schimpl-Neimanns, Bernhard; Rendtel, Ulrich: SAS-, SPSS- und
            STATA-Programme zur Berechnung der Varianz von
            Populationsschaetzern im Mikrozensus. ZUMA-Methodenbericht
            Nr. 2001/04. Mannheim. (URL www.gesis.org/Publikationen/
            Berichte/ZUMA_Methodenberichte/documents/pdfs/tb01_04.pdf )
   5. Projektbeginn: Oktober 1999
   6. Letzte Programmaenderung: 26. Januar 2001
   7. Programmstatus: Getestet mit SAS fuer Windows V8.0 (Windows NT 4.0,
         SP6) und Mikrozensus 1996 (faktisch anonymisierte 70%-
         Substichprobe).
   8. Erforderliche Programmeingaben: SAS-Datensatz basierend auf den
         Rohdaten des Mikrozensus. Das File sollte keine Missing Values
         enthalten.
         Schichtvariablen: EF1 Bundesland, EF708 Gemeindegroessenklasse,
            EF712 Gebaeudegroessenklasse
         Klumpenidentifikation (PSU): EF3 Auswahlbezirk
         Haushaltsidentifikation (HHNR): EF4 Haushaltsnummer
         Die bei anderen als in diesem Beispiel verwendeten Y-Variablen
         und Subpopulationen (Z) sowie insgesamt zu aendernden
         Programmschritte sind mit spitzen Klammern <> gekennzeichnet.
   9. Grobe Programmstruktur:
       A Einzeldaten einlesen und benoetigte Variablen definieren
         Berechnen der Haushaltstotals
       B Haushaltsbezogene Daten weiterverarbeiten (ggf. einlesen)
         Berechnen des Verhaeltnisses R = t_y / t_z
         Berechnen der Hilfsgroesse u=y-R*z
       C Berechnen der PSU Totals und PSU Within Varianzen
         Berechnen der Schicht Totals, der Between Varianz und der
            Summe der gewichteten Within Varianzen
         Summation der Totals und Varianz Terme ueber die Schichten
         Berechnung der Standard Abweichungen
         Berechnung der Varianz unter Annahme der Binomialverteilung
         Berechnung der auszugebenden Kennwerte
         Ausgabe der Kennwerte
   ---------------------------------------------------------------------- */

   /* Teil A:  Daten Personenbezogen  */

LIBNAME <libname> '<Verzeichnis>';
```

```
DATA MZ_pers;
     set <libname.filename>
        (keep = ef1 ef3 ef4 <ef35> <ef506> ef708 ef712);
     PSU=ef3;
     HHNR=ef4;
     schicht = ef1*100 + ef708*10 + ef712;
     y  = <(EF35 eq 1)*(ef506 eq 1)>;
     z  = <(ef506 eq 1)>;
     LABEL  PSU="Auswahlbezirksnummer (EF3)"
            HHNR="lfd. Haushaltsnummer im Auswahlbezirk (EF4)"
            SCHICHT="Bu.land * Gem.groesse * Geb.schicht"
            Y="<Ledige in Privathaushalten (1, sonst 0)>"
            Z="<Bevoelkerung in Privathaushalten (1, sonst 0)>";
     run;

   /* Sortieren falls noetig, sonst ueberspringen */
Proc sort data=MZ_pers;
          by schicht psu hhnr;
     run;

   /*  Berechnen der Haushaltstotals */
Proc Means data=MZ_pers noprint;
     var Y Z;
     by schicht psu hhnr;
     output out=MZ
            sum=y_k z_k;
     run;

   /* Ende Teil A */

   /* Teil B: Daten  Haushaltsbezogen
              Beim Einlesen haushaltsbezogener Daten muessen
              Y und Z als y_k und z_k aggregiert vorliegen  */

    /* Sortieren falls noetig, sonst ueberspringen */
Proc sort data=MZ; by schicht psu hhnr;

     /* Berechnen des Verhaeltnisses R=t_y / t_z  */
Proc means data=MZ;
     var y_k z_k;
     output out=totals
            sum=t_y t_z;

data totals ;
     set totals ;
     R=t_y/t_z;
     t_y=t_y*100/0.7;
     t_z=t_z*100/0.7;

    /* Berechnen der Hilfsgroesse u=y - R*z */
       /* _N_ := match von Haushaltsdaten "MZ" mit "totals":
                 jedem Satz in "MZ" wird das Ergebnis von u=y_k-R*z_k
                 aus "totals" (Datei besteht aus 1 Zeile) zugespielt  */
data MZ;
     set MZ;
     if _N_ eq 1 then set totals;
     u_k=y_k - R * z_k;

   /* Ende Teil B */

   /* Teil C: Ab hier weiter in beiden Faellen  */

   /* Berechnen der PSU Totals und PSU Within Varianzen */
```

```
Proc Means data=MZ noprint;
     Var u_k;
     By schicht psu ;
     output out=psu_data
            sum=psu_u
            Var=psu_var
            N  =psu_n;
     run;

    /* Berechnen der Schicht Totals, der Between Varianz und der Summe der
       gewichteten Within Varianzen */
Data PSU_data;
     set Psu_data;
         /* Einige PSUs sind im FAMZ nur mit E I N E M  Haushalt
            repraesentiert – Missings rekodieren */
     if Nmiss(PSU_Var)=1 then PSU_var=0;
        /* Berechnung n_i*S^2(s_i)   */
     N_Var=PSU_N*PSU_Var;

Proc means data=PSU_data noprint;
     Var PSU_u N_var;
     by Schicht;
     output out=str_data
            sum=STR_u STR_with
            Var=Between
            N  =STR_n ;
     run;

 /* Strata Varianz = 100^2*0.99*n_(I,h)*S^2(n_I,h)/(0.7*0.7) + [= V_betw]
                     100*0.3/(0.7*0.7) * Summe n_i*S^2(s_i)    [= V_with]*/
Data STR_data;
     set STR_data;
     V_betw=10000*0.99*STR_N*Between/(0.7*0.7) ;
     V_with= 100*0.3*STR_with/(0.7*0.7);
     v=v_Betw + V_with ;

   /* Summation der Varianz Terme ueber die Strata */
Proc Means data=STR_data  noprint;
     Var  V_betw V_with V;
     output out=ergebnis
            sum =  V_betw V_with V;
     run;

   /* Berechnung der Standard Abweichungen und Ausdrucken der Ergebnisse */
data ergebnis ;
    merge ergebnis totals;
       /* Division der Varianzen durch t_z^2 */
    V=V/(t_z*t_z);
    V_betw=V_betw/(t_z*t_z);
    V_with=V_with/(t_z*t_z);
    sig_V=sqrt(V);
    sig_B=sqrt(V_betw);
    sig_W=sqrt(V_with);
    rel=sig_V/r;
    n_sub=t_z*0.7/100;    /* Subpopulation n */
      /* BINOMIALVARIANZ NUR FUER ANTEILSWERTE */
    sig_bin=sqrt( (r*(1-r))/(n_sub-1)); /* Binomialverteilung */
    deft=sig_v/sig_bin;
    r=r*100;              /* Ausgabe in Prozent */
    sig_v=sig_v*100;
    sig_b=sig_b*100;
    sig_w=sig_w*100;
    rel=rel*100;
```

```
    t_y=t_y/1000;         /* Ausgabe in 1000 */
    t_z=t_Z/1000;
    Label r="Ratio t_y/t_z (in %)"
          sig_v="Std.Fehler (R) Insgesamt (in %)"
          sig_b="Std.Fehler Between-Teil (in %)"
          sig_w="Std.Fehler Within-Teil (in %)"
          rel="relativer Std.Fehler (in %)"
          deft="Design-Effekt Faktor"
          t_y="Total y (in 1000)"
          t_z="Total z - Subpopulation (in 1000)";

Proc print data=ergebnis;
    var r sig_v sig_b sig_w rel deft t_y t_z ;
    run;

/* ========= PROGRAMMAENDERUNGEN BEI MITTELWERTEN ====================
 Beispiel: Merkmal Y "Heiratsalter von verheirateten und mit dem Partner
                     zusammenlebenden Frauen ..."
           Subpopulation Z "verheiratete, mit dem Partner zusammenlebende
*                     Frauen, Bevoelkerung am Hauptwohnsitz, gueltige
*                     Angaben zum Heiratsjahr und Heiratsjahr>=1925,
*                     Geburtsjahr>=1901" .
 (...)
DATA MZ_pers ;
     set <libname.filename> ;
        (keep = ef1 ef3 ef4 <ef32> <ef33> <ef35> <ef36> <ef505> <ef575>
                ef708 ef712);
     (...)
     z =<(ef32 eq 2 & ef505 ge 1 & ef505 le 2 & ef35 eq 2 &
         ef575 ge 1 & ef575 le 3 & ef33 ge 1901 &
         ef36 ge 1925 & ef36 le 1996)> ;
     y =<z*(ef36-ef33)>;
     LABEL  (...)
            Y="<Heiratsalter v. Frauen ...>"
            Z="<verh. Frauen, Bevoelkerung am Hauptwohnsitz ...>" ;
    (...)
data ergebnis ;
     merge ergebnis totals;
     sig_V=sqrt(V);
     sig_B=sqrt(V_betw);
     sig_W=sqrt(V_with);
     rel=100*sig_V/r;
Proc print data=ergebnis;
     var r sig_v sig_b sig_w rel t_y t_z;
========= PROGRAMMAENDERUNGEN BEI MITTELWERTEN ======================== */
```

```
/* ------------------------- VarMZ_A.SAS --------------------------------
   1. Programmname: VarMZ_A.SAS (URL www.gesis.org/Dauerbeobachtung/
         Mikrodaten/mikrodaten_tools/Varianz/VarMZ_A.SAS )
   2. Programmautoren: Ulrich Rendtel (rendtel@em.uni-frankfurt.de)
         Bernhard Schimpl-Neimanns (schimpl-neimanns@zuma-mannheim.de)
   3. Zweck des Programms: Berechnung der Varianz des Regressionsschaetzers
         bei Randanpassung von Mikrozensus-Fallzahlen an die
         Bevoelkerungsfortschreibung im faktisch anonymisierten
         Mikrozensus (FAMZ) 1996.
         Hier am Beispiel des Merkmals Y "Ledige" (EF35=1) fuer die
         Subpopulation Z "Bevoelkerung in Privathaushalten" (EF506=1)
   4. Weiterfuehrende Aufgabenbeschreibungen:
         Rendtel, Ulrich; Schimpl-Neimanns, Bernhard: Varianzschaetzungen
            fuer den faktisch anonymisierten Mikrozensus. In: Jahrbuecher
            fuer Nationaloekonomie und Statistik, 220/6, 2000, S. 759-776.
         Rendtel, Ulrich; Schimpl-Neimanns, Bernhard: Die Berechnung der
            Varianz von Populationsschaetzern im Scientific Use File des
            Mikrozensus. In: ZUMA-Nachrichten Nr. 48, 2001, S. 85-116
            (URL www.gesis.org/Publikationen/Zeitschriften/
            ZUMA_Nachrichten/documents/pdfs/zn48_10-bernhard.pdf )
         Schimpl-Neimanns, Bernhard; Rendtel, Ulrich: SAS-, SPSS- und
            STATA-Programme zur Berechnung der Varianz von
            Populationsschaetzern im Mikrozensus. ZUMA-Methodenbericht
            Nr. 2001/04. Mannheim. (URL www.gesis.org/Publikationen/
            Berichte/ZUMA_Methodenberichte/documents/pdfs/tb01_04.pdf )
   5. Projektbeginn: Oktober 1999
   6. Letzte Programmaenderung: 26. Januar 2001
   7. Programmstatus: Getestet mit SAS fuer Windows V8.0 (Windows NT 4.0,
         SP6) und Mikrozensus 1996 (faktisch anonymisierte 70%-
         Substichprobe).
   8. Erforderliche Programmeingaben: SAS-Datensatz basierend auf den
         Rohdaten des Mikrozensus. Das File sollte keine Missing Values
         enthalten.
         Schichtvariablen: EF1 Bundesland, EF708 Gemeindegroessenklasse,
            EF712 Gebaeudegroessenklasse
         Klumpenidentifikation (PSU): EF3 Auswahlbezirk
         Haushaltsidentifikation (HHNR): EF4 Haushaltsnummer
         Abgrenzung der Anpassungsklassen (ANP): EF32 Geschlecht,
            EF52 Staatsangehoerigkeit, EF127 Stellung im Beruf
         Gruppendefinition (GRUPPE): EF1 Bundesland, ANP
         Die bei anderen als in diesem Beispiel verwendeten Y-Variablen
         und Subpopulationen (Z) sowie insgesamt zu aendernden
         Programmschritte sind mit spitzen Klammern <> gekennzeichnet.
   9. Grobe Programmstruktur:
       Einzeldaten einlesen und benoetigte Variablen definieren
       Berechnen des Regressionskoeffizienten B^
       Ausgabe des SOLL/IST Vergleichs fuer die Gruppen und des
          gewichteten Gesamt-Totals ueber die Gruppen
       Berechnen der Hilfsgroesse u=g * (y - B^*1)
       Berechnen der Haushaltstotals der Hilfsgroesse
       Haushaltsbezogene Daten weiterverarbeiten
       Berechnen des Verhaeltnisses R=t_y / t_z
       Berechnen der Hilfsgroesse u=y - R*z
       Berechnen der PSU Totals und PSU Within Varianzen
       Berechnen der Schicht Totals, der Between Varianz und der
          Summe der gewichteten Within Varianzen
       Summation der Totals und Varianz Terme ueber die Schichten
       Berechnung der Standard Abweichungen
       Berechnung der auszugebenden Kennwerte
       Ausgabe der Kennwerte
   ------------------------------------------------------------------- */

LIBNAME <libname> '<verzeichnis>' ;
```

```
DATA MZ_pers ;
     set <libname.filename>
       (keep = ef1 ef3 ef4 ef32 <ef35> ef52 ef127 <ef506>
               ef708 ef712 <ef750>);
     PSU=ef3;
     HHNR=ef4;
     schicht = ef1*100 + ef708*10 + ef712;
     soll_ist=<ef750>;
     Y  =<(ef35 eq 1)*(ef506 eq 1)>;
     Y_W=y*soll_ist;
     Z  =<(ef506 eq 1)>;
     if (ef52 eq 1 AND ef32 eq 1 AND ef127 ne 9 AND ef127 ne 10)
         then anp=1;
         else if (ef52 eq 1 AND ef32 eq 2) then anp=2;
         else if (ef52 ne 1 AND ef32 eq 1) then anp=3;
         else if (ef52 ne 1 AND ef32 eq 2) then anp=4;
         else if (ef52 eq 1 AND ef32 eq 1 AND ef127 eq 9) then anp=5;
         else if (ef52 eq 1 AND ef32 eq 1 AND ef127 eq 10) then anp=6;
         * ANPASSUNGSKLASSEN: 1=Deutsche Maenner, 2=Deutsche Frauen ;
         *      3=Auslaendische Maenner, 4=Auslaendische Frauen ;
         *      5=Zeit-/Berufssoldaten 6=Wehrdienstleistende ;
     GRUPPE=ef1*10+anp;
     LABEL  PSU="Auswahlbezirksnummer (EF3)"
            HHNR="lfd. Haushaltsnummer im Auswahlbezirk (EF4)"
            SCHICHT="Bu.land * Gem.groesse * Geb.schicht"
            Y="<Ledige in Privathaushalten (1, sonst 0)>"
            Y_W="Y - gewichtete Beobachtung "
            Z="<Bevoelkerung in Privathaushalten (1, sonst 0)>"
            ANP="Anpassungsklassen"
            GRUPPE="Hochrechnungsgruppen";
     run;

   /* Sortieren nach Gruppen */
Proc sort data=MZ_pers; by gruppe; run;

   /* Berechnen des Regressionskoeffizienten B_dach=t_y_g / t_x_g  */
Proc means data=MZ_pers noprint;
     var y y_w Soll_ist;
     by gruppe;
     output out=Reg_koef
           Sum =t_y_g t_y_w t_s_i
           N   =t_x_g t_x_w;

data Reg_koef ;
     set Reg_koef ;
     B_dach =t_y_g/t_x_g;
     t_y_w=t_y_w*100/0.7;
     t_ist =t_x_w*100/0.7;
     t_soll=t_ist *t_s_i/t_x_g;
     run;

     /* Ausgabe des Soll/Ist Vergleichs fuer die einzelnen Gruppen */
Proc Print data=Reg_koef ;
     var gruppe B_dach t_ist t_soll t_y_w;
     run;

     /* Berechnung des gewichteten Gesamt-Totals ueber die Gruppen */
Proc means data=Reg_koef noprint;
     var t_y_w ;
     output out=erg_est
            sum= T_w ;
     run;
```

```
    /* Berechnen der Hilfsgroesse u=g*(y - B_dach*1) mit g=Soll_Ist  */
data MZ_pers;
     Merge  MZ_pers reg_koef;
     by gruppe ;
     u=soll_ist*(y-B_dach);

    /* Sortieren nach Schicht, PSU und Haushalts-Nr. */
Proc sort data=MZ_pers  ;
     by schicht psu hhnr;

     /* Berechnung der Haushaltstotals von u */
Proc means data=MZ_pers noprint;
     var u;
     by schicht psu hhnr;
     output out=MZ_h
            sum=u_k;


   /* Berechnen der PSU Totals und PSU Within Varianzen */
Proc Means data=MZ_h noprint;
     var u_k;
     by schicht psu ;
     output out=psu_data
           sum=psu_u
           Var=psu_var
           N  =psu_n;
     run;

    /* Berechnen der Schicht Totals, der Between Varianz und der
       Summe der gewichteten Within Varianzen */

Data PSU_data;
     set Psu_data;
     /* Einige PSUs sind im FAMZ nur mit E I N E M  Haushalt repraesentiert
        - Rekodieren der Missing Values */
     if Nmiss(PSU_Var)=1 then PSU_var=0;
     /* Berechnung n_i*S^2(s_i)   */
     N_Var=PSU_N*PSU_Var;

Proc means data=PSU_data noprint;
     Var PSU_u N_var;
     by Schicht;
     output out=str_data
            sum=STR_u STR_with
            Var=Between
            N  =STR_n ;
     run;

   /* Strata Varianz=100^2*0.99*n_(I,h)*S^2(n_I,h)/(0.7*0.7)  [= V_betw]
                     + 100*0.3/(0.7*0.7) * Summe n_i*S^2(s_i) [= V_with] */
Data STR_data;
     set STR_data;
     V_betw=10000*0.99*STR_N*Between/(0.7*0.7) ;
     V_with= 100*0.3*STR_with/(0.7*0.7);
     v=v_Betw + V_with ;

   /* Summation der Varianz Terme ueber die Strata */
Proc Means data=STR_data  noprint;
     var  V_betw V_with V;
     output out=ergebnis
            sum=V_betw V_with V;
     run;
```

```
   /* Berechnung der Standard Abweichungen und Ausdrucken der Ergebnisse */
data ergebnis ;
     merge ergebnis erg_est;
     t_w = t_w/1000;
     sig_V=sqrt(V)/1000;
     sig_B=sqrt(V_betw)/1000;
     sig_W=sqrt(V_with)/1000;
     rel=100*sig_V/t_w;
     LABEL t_w="gewichtetes Total (in 1000)"
           sig_v="Std.Fehler (in 1000)"
           sig_b="Std.Fehler Between-Teil (in 1000)"
           sig_w="Std.Fehler Within-Teil (in 1000)"
           rel = "relativer Std.Fehler (in %)";
     run;

Proc print data=ergebnis;
     var t_w sig_v sig_b sig_w rel;
     run;
```

```
* ------------------------- VarMZ_T.SPS ------------------------------.
* 1. Programmname: VarMZ_T.SPS (URL www.gesis.org/Dauerbeobachtung/    .
*       Mikrodaten/mikrodaten_tools/Varianz/VarMZ_T.SPS )              .
* 2. Programmautoren: Ulrich Rendtel (rendtel@em.uni-frankfurt.de),    .
*       Bernhard Schimpl-Neimanns (schimpl-neimanns@zuma-mannheim.de)  .
* 3. Zweck des Programms: Berechnung der Varianz des Pi-Schaetzers fuer.
*       ein Merkmal Y im faktisch anonymisierten Mikrozensus 1996      .
*       Hier am Beispiel des Merkmals Y "Ledige" (EF35=1) fuer die     .
*          Subpopulation Z "Bevoelkerung in Privathaushalten" (EF506=1).
* 4. Weiterfuehrende Aufgabenbeschreibungen:                           .
*       Rendtel, Ulrich; Schimpl-Neimanns, Bernhard:                   .
*          Varianzschaetzungen fuer den faktisch anonymisierten        .
*          Mikrozensus. In: Jahrbuecher fuer Nationaloekonomie und     .
*          Statistik, 220/6, 2000, S. 759-776                          .
*       Rendtel, Ulrich; Schimpl-Neimanns, Bernhard: Die Berechnung der.
*          Varianz von Populationsschaetzern im Scientific Use File des.
*          Mikrozensus. In: ZUMA-Nachrichten Nr. 48, 2001, S. 85-116   .
*          (URL www.gesis.org/Publikationen/Zeitschriften/             .
*          ZUMA_Nachrichten/documents/pdfs/zn48_10-bernhard.pdf )      .
*       Schimpl-Neimanns, Bernhard; Rendtel, Ulrich: SAS-, SPSS- und   .
*          STATA-Programme zur Berechnung der Varianz von              .
*          Populationsschaetzern im Mikrozensus. ZUMA-Methodenbericht  .
*          Nr. 2001/04. Mannheim. (URL www.gesis.org/Publikationen/    .
*          Berichte/ZUMA_Methodenberichte/documents/pdfs/tb01_04.pdf ) .
* 5. Projektbeginn: Oktober 1999                                       .
* 6. Letzte Programmaenderung: 26. Januar 2001                         .
* 7. Programmstatus: Getestet mit SPSS for Windows, Release 10.0.5 .
*       (Windows NT 4.0, SP6) und Mikrozensus 1996 (faktisch .
*        anonymisierte 70%-Substichprobe).
* 8. Erforderliche Programmeingaben: SPSS-Datensatz basierend auf den  .
*       Rohdaten des Mikrozensus ohne Missing Value Deklaration        .
*       Schichtvariablen: EF1 Bundesland, EF708 Gemeindegroessenklasse,.
*          EF712 Gebaeudegroessenklasse                                .
*       Klumpenidentifikation (PSU): EF3 Auswahlbezirk                 .
*       Haushaltsidentifikation (HHNR): EF4 Haushaltsnummer            .
*       Die bei anderen als in diesem Beispiel verwendeten Y-Variablen .
*          und Subpopulationen (Z) sowie insgesamt zu aendernden       .
*          Programmschritte sind mit spitzen Klammern <> gekennzeichnet.
* 9. Grobe Programmstruktur:                                           .
*      A Einzeldaten einlesen und benoetigte Variablen definieren      .
*        Berechnen der Haushaltstotals                                 .
*      B Haushaltsbezogene Daten weiterverarbeiten (ggf. einlesen)     .
*      C Berechnen der PSU Totals und PSU Within Varianzen             .
*        Berechnen der Schicht Totals, der Between Varianz und der     .
*           Summe der gewichteten Within Varianzen                     .
*        Summation der Totals und Varianz Terme ueber die Schichten    .
*        Berechnung der Standard Abweichungen                          .
*        Berechnung der Varianz unter Annahme der Binomialverteilung   .
*        Berechnung der auszugebenden Kennwerte                        .
*        Ausgabe der Kennwerte                                         .
* --------------------------------------------------------------------.

set width 80.
set length none.
set compression=on.
set header=no.
set mxwarns=300000.
set blanks=sysmis.
set mxmemory=2097151.

* Teil A: Daten Personenbezogen .
```

```
get file '<filename>'
   /keep ef1 ef3 ef4 <ef35> <ef506> ef708 ef712
   /rename (ef3 ef4 = psu hhnr).
missing values all().
weight off.

* Subpopulation definieren.
compute z=0.
if (<ef506=1>) z=1.
variable label z '<Bevoelkerung in Privathaushalten>'.

* Interessierendes Merkmal definieren.
compute y=0.
if (<z=1 & ef35=1>) y=1.
string ytext (A80).
variable label ytext 'Y-Variable - Z-Subpopulation'.
compute ytext = '<Ledige in Privathaushalten (ef35=1 & ef506=1)>'.

* Protokollierung bis Ausgabe der Kennwerte unterdruecken .
set messages=none.
set printback=none.

* Schichtvariable bilden: .
compute schicht = ef1*100 + ef708*10 + ef712.
variable label schicht 'Bu.land | Gem.groesse | Gebaeudeschicht'.
formats schicht (f4).

* Berechnen der Haushaltstotals.
sort cases by schicht psu hhnr.
aggregate /outfile=*
          /presorted
          /break = schicht psu hhnr
          /y_k 'Merkmalswert y im Haushalt k' = sum(y)
          /z_k 'Anzahl Personen im Haushalt k' = sum(z)
          /ytext = first(ytext).

* Teil B: Daten Haushaltsbezogen .
*         Beim Einlesen haushaltsbezogener Daten muessen .
*         Y und Z als y_k und z_k aggregiert vorliegen.

* Sortieren falls noetig, sonst ueberspringen .
sort cases by schicht psu hhnr.

* Teil C: Ab hier weiter in beiden Faellen .

*   Berechnung der PSU Totals und PSU Within Varianzen .
aggregate /outfile=*
          /presorted
          /break = schicht psu
          /psu_y 'Merkmalswert y in PSU i' = sum(y_k)
          /psu_n 'Anzahl Haushalte in PSU i' = n
          /psu_var 'PSU Within Varianz' = sd(y_k)
          /psu_pers 'Anzahl Personen in PSU i' = sum(z_k)
          /ytext = first(ytext).

* psu_var: Standardabweichung (SD) => Varianz.
compute psu_var=psu_var * psu_var.
* Einige PSUs sind im FAMZ nur mit E I N E M  Haushalt repraesentiert .
*   Missings rekodieren .
recode psu_var (missing=0).

* n_var: Berechnung n_i*S^2(s_i)   .
compute n_var=psu_n*psu_var.
```

```
* Fuer Berechnung der Schicht Totals, Between Varianz und Summe .
* der gewichteten Within Varianzen Schicht-File erzeugen .
aggregate /outfile=*
          /presorted
          /break = schicht
          /str_y 'Merkmalswert y in Schicht h' = sum(psu_y)
          /str_with 'Within Varianz' = sum(n_var)
          /str_n 'Anzahl PSUs in Schicht h' = n
          /between 'Between Varianz' = sd(psu_y)
          /str_pers 'Anzahl Personen in Schicht h' = sum(psu_pers)
          /ytext = first(ytext).

* between: Std.abweichung (SD) => Varianz .
compute between=between*between.
recode between (missing=0).

compute total=100*str_y/0.7.

* Strata Varianz = 100^2*0.99*n_(I,h)*S^2(n_I,h)          [= V_betw].
*                 +100*0.3/(0.7*0.7) * Summe n_i*S^2(s_i) [= V_with].
compute v_betw=100*100*0.99*str_n*between/(0.7*0.7).
compute v_with=100*0.3/(0.7*0.7)*str_with.
compute v=v_betw+v_with.

* Summation der Totals und Varianz Terme ueber die Schichten .
compute insges=1.
aggregate outfile=*
          /presorted
          /break = insges
          /total 'Total (t^)' = sum(total)
          /v 'Varianz V(t^)'= sum(v)
          /v_with 'Varianz Within-Teil' = sum(v_with)
          /v_betw 'Varianz Between-Teil' = sum(v_betw)
          /y_insg 'Summe Y=1' = sum(str_y)
          /n_insg 'Summe Y=0,1' = sum(str_pers)
          /ytext = first(ytext).

* Berechnung der Standardabweichungen.
compute sig_V=sqrt(v).
compute sig_W=sqrt(v_with).
compute sig_B=sqrt(v_betw).

* relativer Std.fehler.
compute rel=sig_V/total.

* Relativer Standardfehler nach dem Binomialansatz (Rel_Bin) .
*   v^(bin) .
compute p_dach = y_insg / n_insg.
compute Rel_Bin = sqrt( 0.99 * (1-p_dach) / (p_dach*(n_insg - 1)) ).

* Design-Effekt Faktor: rel.s.e. FAMZ/rel.s.e. Binomialansatz .
compute deft = rel/rel_bin.

comment *** Berechnung der Werte in den entsprechenden Einheiten ***.
compute p_dach=p_dach*100.
compute rel=rel*100.
compute rel_bin=rel_bin*100.
compute total=total/1000.
compute sig_V=sig_V/1000.
compute sig_B=sig_B/1000.
compute sig_W=sig_W/1000.
```

```
variable label total   'Total (t^) (in 1000)'.
variable label sig_v   'Std.Fehler (t^) (in 1000)'.
variable label sig_W   'Std.Fehler Within-Teil (in 1000)'.
variable label sig_B   'Std.Fehler Between-Teil (in 1000)'.
variable label rel     'Relativer Std.fehler (sig_V/Total) (in %)'.
variable label p_dach  'Anteil Y in Subpopulation Z (in %)'.
variable label rel_bin 'Rel.Std.fehler nach Binomialansatz (in %)'.
variable label deft    'Design-Effekt Faktor (rel_/rel_bin)'.

formats Total v v_betw v_with sig_V sig_B sig_W
      p_dach  rel rel_bin deft (f16.4).
formats ytext (a80).

* Ausgabe der Kennwerte .
set messages=listing.
set printback=listing.
summarize
  /tables= ytext total sig_V sig_B sig_W rel deft rel_bin
  /format=nolist /cells=first.
```

```
* ------------------------- VarMZ_R.SPS -------------------------------.
* 1. Programmname: VarMZ_R.SPS (URL www.gesis.org/Dauerbeobachtung/ .
*       Mikrodaten/mikrodaten_tools/Varianz/VarMZ_R.SPS ) .
* 2. Programmautoren: Ulrich Rendtel (rendtel@em.uni-frankfurt.de) .
*       Bernhard Schimpl-Neimanns (schimpl-neimanns@zuma-mannheim.de) .
* 3. Zweck des Programms: Berechnung der Varianz des Pi-Schaetzers fuer .
*       das Verhaeltnis R zweier Totals, t_y und t_z, im faktisch .
*       anonymisierten Mikrozensus (FAMZ) 1996.
*       Hier am Beispiel des Merkmals Y "Ledige" (EF35=1) fuer die .
*       Subpopulation Z "Bevoelkerung in Privathaushalten" (EF506=1) .
*       Das Programm kann auch zur Berechnung der Varianz des .
*       arithmetischen Mittelwerts der Variablen Y fuer die Subpopulation .
*       Z verwendet werden. (Ein Beispiel befindet sich am Ende des .
*       Programms.) .
* 4. Weiterfuehrende Aufgabenbeschreibungen: .
*       Rendtel, Ulrich; Schimpl-Neimanns, Bernhard: Varianzschaetzungen .
*          fuer den faktisch anonymisierten Mikrozensus. In: Jahrbuecher .
*          fuer Nationaloekonomie und Statistik, 220/6, 2000, S. 759-776 .
*       Rendtel, Ulrich; Schimpl-Neimanns, Bernhard: Die Berechnung der .
*          Varianz von Populationsschaetzern im Scientific Use File des .
*          Mikrozensus. In: ZUMA-Nachrichten Nr. 48, 2001, S. 85-116 .
*          (URL www.gesis.org/Publikationen/Zeitschriften/ .
*          ZUMA_Nachrichten/documents/pdfs/zn48_10-bernhard.pdf ) .
*       Schimpl-Neimanns, Bernhard; Rendtel, Ulrich: SAS-, SPSS- und     .
*          STATA-Programme zur Berechnung der Varianz von .
*          Populationsschaetzern im Mikrozensus. ZUMA-Methodenbericht .
*          Nr. 2001/04. Mannheim. (URL www.gesis.org/Publikationen/ .
*          Berichte/ZUMA_Methodenberichte/documents/pdfs/tb01_04.pdf ).
* 5. Projektbeginn: Oktober 1999 .
* 6. Letzte Programmaenderung: 26. Januar 2001 .
* 7. Programmstatus: Getestet mit SPSS for Windows, Release 10.0.5, .
*       (Windows NT 4.0, SP6) und Mikrozensus 1996 (faktisch .
*       anonymisierte 70%-Substichprobe).
* 8. Erforderliche Programmeingaben: SPSS-Datensatz basierend auf den .
*    Rohdaten des Mikrozensus ohne Missing Value Deklaration.
*    Schichtvariablen: EF1 Bundesland, EF708 Gemeindegroessenklasse, .
*       EF712 Gebaeudegroessenklasse
*    Klumpenidentifikation (PSU): EF3 Auswahlbezirk .
*    Haushaltsidentifikation (HHNR): EF4 Haushaltsnummer .
*    Die bei anderen als in diesem Beispiel verwendeten Y-Variablen  .
*       und Subpopulationen (Z) sowie insgesamt zu aendernden  .
*       Programmschritte sind mit spitzen Klammern <> gekennzeichnet.
* 9. Grobe Programmstruktur: .
*    A Einzeldaten einlesen und benoetigte Variablen definieren .
*      Berechnen der Haushaltstotals .
*    B Haushaltsbezogene Daten weiterverarbeiten (ggf. einlesen) .
*      Berechnen des Verhaeltnisses R = t_y / t_z .
*      Berechnen der Hilfsgroesse u=y-R*z .
*    C Berechnen der PSU Totals und PSU Within Varianzen .
*      Berechnen der Schicht Totals, der Between Varianz und der  .
*         Summe der gewichteten Within Varianzen .
*         Summation der Totals und Varianz Terme ueber die Schichten .
*         Berechnung der Standard Abweichungen .
*         Berechnung der Varianz unter Annahme der Binomialverteilung  .
*         Berechnung der auszugebenden Kennwerte .
*         Ausgabe der Kennwerte .
*---------------------------------------------------------------------.
set width 80.
set length none.
set compression=on.
set header=no.
set mxwarns=300000.
set blanks=sysmis.
```

```
set mxmemory=2097151.

* Teil A: Daten Personenbezogen .

get file '<filename>'
   /keep ef1 ef3 ef4 <ef35> <ef506> ef708 ef712
   /rename (ef3 ef4 = psu hhnr).
missing values all().
weight off.

* Subpopulation definieren.
compute z=0.
if (<ef506=1>) z=1.
variable label z '<Bevoelkerung in Privathaushalten>'.

* Interessierendes Merkmal definieren.
compute y=0.
if (<z=1 & ef35=1>) y=1.
string ytext (A80).
variable label ytext 'Y-Variable - Z-Subpopulation'.
compute ytext = '<Ledige in Privathaushalten (ef35=1 & ef506=1)>'.

* Protokollierung bis Ausgabe der Kennwerte unterdruecken .
set messages=none.
set printback=none.

* Schichtvariable bilden: .
compute schicht = ef1*100 + ef708*10 + ef712.
variable label schicht 'Bu.land | Gem.groesse | Gebaeudeschicht'.
formats schicht (f4).

* Berechnen des Verhaeltnisses R=t_y/t_z.

*  1. Schritt: Fallzahlsummen fuer gesamte Stichprobe schreiben.
sort cases by schicht psu hhnr.
compute eins=1.
temporary.
aggregate /outfile='<temp.sav>'
          /presorted
          /break = eins
          /t_y = sum(y)
          /t_z = sum(z)
          /n_insg 'Stichprobengroesse' = n
          /n_sub 'Subpopulation n' = sum(z).
          /ytext = first(ytext).

* 2. Schritt: Fallzahlsummen den Einzeldaten zuspielen.
match files /file *
            /table '<temp.sav>'
            /by eins.
execute.

* 3. Schritt: Berechnen des Verhaeltnisses R .
compute r = t_y/t_z.

* 4. Schritt: Berechnen der Hilfsgroesse u = y - R * z ("Ratio-Residual").
compute u = y - r*z.
execute.

* ENDE TEIL A: PERSONENBEZOGENE DATEN .

subtitle 'Teil B: Daten Haushaltsbezogen '.
```

```
*    Beim Einlesen haushaltsbezogener Daten muessen .
*    Y und Z als y_k und z_k aggregiert vorliegen .

* sortieren falls noetig, sonst ueberspringen .
sort cases by schicht psu hhnr.

aggregate /outfile=*
          /presorted
          /break = schicht psu hhnr
          /y_k 'Merkmalswert y im Haushalt' = sum(y)
          /z_k 'Merkmalswert z im Haushalt' = sum(z)
          /u_k 'Ratio-Residual im Haushalt' = sum(u)
          /n_k 'Haushaltsgroesse' = n
          /n_insg 'Stichprobengroesse' = first(n_insg)
          /n_sub 'Subpopulation n' = first(n_sub)
          /t_y = first(t_y)
          /t_z = first(t_z)
          /r = first(r)
          /ytext = first(ytext).

* alternative Berechnung der Hilfsgroesse u=y-R*z ("Ratio-Residual").
*             auf Haushaltsebene:  .
* compute u_k = y_k - (t_y/t_z)*z_k .


* Teil C: Berechnung der PSU Totals und PSU Within Varianzen .

aggregate /outfile=*
          /presorted
          /break = schicht psu
          /psu_u 'Hilfsgroesse u in PSU i' = sum(u_k)
          /psu_n 'Anzahl Haushalte in PSU i' = n
          /psu_var 'PSU Within Varianz' = sd(u_k)
          /psu_pers 'Anzahl Personen in PSU i' = sum(n_k)
          /t_y = first(t_y)
          /t_z = first(t_z)
          /r = first(r)
          /n_insg=first(n_insg)
          /n_sub=first(n_sub)
          /ytext = first(ytext).

* Einige PSUs sind im FAMZ nur mit E I N E M  Haushalt repraesentiert: .
*        Missings rekodieren.
recode psu_var (missing=0).
* psu_var: Standardabweichung (SD) => Varianz.
compute psu_var=psu_var * psu_var.

* n_var: Berechnung n_i*S^2(s_i)   .
compute n_var=psu_n*psu_var.

* Berechnen der Schicht Totals, Between Varianz und Summe .
*   der gewichteten Within Varianzen .
aggregate /outfile=*
          /presorted
          /break = schicht
          /str_u 'Merkmalswert u in Schicht h' = sum(psu_u)
          /str_with 'Within Varianz (u)' = sum(n_var)
          /str_n 'Anzahl PSUs in Schicht h' = n
          /between 'Between Varianz' = sd(psu_u)
          /str_pers 'Anzahl Personen in Schicht h' = sum(psu_pers)
          /t_y = first(t_y)
          /t_z = first(t_z)
          /r = first(r)
```

```
          /n_insg=first(n_insg)
          /n_sub=first(n_sub)
          /ytext = first(ytext).

* between: Std.abweichung (SD) => Varianz .
recode between (missing=0).
compute between=between*between.

* Schicht Varianz = 100^2*0.99*n_(I,h)*S^2(n_I,h)/0.7*0.7) + [=V_betw] .
*                   100*0.3/(0.7*0.7) * Summe n_I*S^2(s_I)   [=V_with] .
compute v_betw=100*100*0.99*str_n*between/(0.7*0.7).
compute v_with=100*0.3*str_with/(0.7*0.7).
compute v=v_betw+v_with.

* Summation der Totals und Varianz Terme ueber die Schichten .
compute insges=1.
aggregate outfile=*
          /presorted
          /break = insges
          /t_y 'Total (t^_y)' = first(t_y)
          /t_z 'Total (t^_z)' = first(t_z)
          /r 'Ratio R = t^_y / t^_z' = first(r)
          /v 'Varianz V(t^)'= sum(v)
          /v_with 'Varianz Within-Teil' = sum(v_with)
          /v_betw 'Varianz Between-Teil' = sum(v_betw)
          /ytext = first(ytext)
          /n_insg=first(n_insg)
          /n_sub=first(n_sub).

* freie Hochrechnung (100/0,7) der Totals .
compute t_y=t_y*100/0.7.
variable label t_y 'Total Merkmal Y'.
compute t_z=t_z*100/0.7.
variable label t_z 'Total Merkmal Z'.


* Berechnung der Standardfehler .
*    abschliessend Division der Varianzen durch t_z^2 .
compute v=v/(t_z*t_z).
compute var_b=var_b/(t_z*t_z).
compute var_w=var_w/(t_z*t_z).
compute sig_V=sqrt(v).
compute sig_B=sqrt(var_b).
compute sig_W=sqrt(var_w).
compute rel=100*sig_V/r.
compute rel=sig_V/r.

variable label sig_V 'Std.Fehler (R)'.
variable label sig_B 'Std.Fehler (R) Between-Teil'.
variable label sig_W 'Std.Fehler (R) Within-Teil'.
variable label rel 'relativer Std.Fehler (in %)'.

* Ausgabe der Kennwerte in 1000 oder Prozent vorbereiten .
compute r=r*100.
compute sig_v=sig_v*100.
compute sig_B=sig_B*100.
compute sig_W=sig_w*100.
compute rel=rel*100.
compute t_z=t_z/1000.
compute t_y=t_y/1000.

* === NUR FUER ANTEILSWERTE === .
* sig_bin := Standardfehler einstufige ungeschichtete Auswahl .
```

```
compute p=r/100.
compute sig_bin = 100*sqrt( (p * (1-p)) / (n_sub - 1) ).
variable label sig_bin 'Std.Fehler einfache Binomialverteilung (in %)'.
compute deft = sig_v/sig_bin.
variable label deft 'Design-Effekt Faktor des Std.fehlers'.

formats t_z t_y r sig_v sig_B sig_W rel sig_bin deft n_insg n_sub (f10.6).
formats ytext (A80).

* Ausgabe der Kennwerte .
set messages=listing.
set printback=listing.
SUMMARIZE
  /TABLES= ytext t_z t_y r sig_v sig_B sig_W rel
           sig_bin deft
  /FORMAT=NOLIST /CELLS=FIRST .

* ============= PROGRAMMAENDERUNGEN BEI MITTELWERTEN ==================== .
* Beispiel: Merkmal Y "Heiratsalter von verheirateten und mit dem Partner .
*                      zusammenlebenden Frauen ..." .
*           Subpopulation Z "verheiratete, mit dem Partner zusammenlebende.
*                      Frauen, Bevoelkerung am Hauptwohnsitz, gueltige .
*                      Angaben zum Heiratsjahr und Heiratsjahr>=1925, .
*                      Geburtsjahr>=1901" .
* (...) .
* get file <filename>'                                                  .
*   /keep ef1 ef3 ef4 <ef32 ef33 ef35 ef36 ef505 ef575> ef708 ef712  .
*   /rename (ef3 ef4 = psu hhnr)                                        .
* (...) .
* compute z=0.
* if (<ef32=2 & ef505>=1 & ef505<=2 & ef35=2 & ef575>=1 & ef575<=3 &
    ef33>=1901 & ef36>=1925 & ef36<=1996>) z=1.
* variable label z "<Subpop.: weibliche Bevoelkerung am Hauptwohns. etc.>".
* Y-Variable definieren: <Heiratsalter = .
*           Eheschliessungsjahr (EF36) - Geburtsjahr (EF33)> .
* compute y=0.
* Compute y=<z*(ef36-ef33)> .
* variable label y "<Heiratsalter fuer Subpopulation>".
* string ytext (A80).
* variable label ytext 'Y-Variable - Subpopulation'.
* compute ytext = '<Heiratsalter verh. Frauen, Bev. am Hauptwohnsitz>'.
* (...) .
* DIESE ZEILEN (*compute ...) LOESCHEN BZW. AUSKOMMENTIEREN: .
*       KENNWERTE UNVERAENDERT AUSGEBEN .
*       Ausgabe der Kennwerte in 1000 oder Prozent vorbereiten .
* *compute r=r*100.
* *compute sig_v=sig_v*100.
* *compute sig_B=sig_B*100.
* *compute sig_W=sig_w*100.
*
* variable label sig_V 'Std.Fehler (R) '.
* variable label sig_B 'Std.Fehler (R) Between-Teil'.
* variable label sig_W 'Std.Fehler (R) Within-Teil'.
* variable label rel 'relativer Std.Fehler (in %)'.
* formats t_z t_y r sig_v sig_B sig_W rel n_insg n_sub (f12.6).
* formats ytext (A80).
* Ausgabe der Kennwerte .
* set messages=listing.
* set printback=listing.
* SUMMARIZE
  /TABLES= ytext r sig_v sig_B sig_W rel
  /FORMAT=NOLIST /CELLS=FIRST .
* =============PROGRAMMAENDERUNGEN BEI MITTELWERTEN ==================== .
```

```
* ------------------------- VarMZ_A.SPS ------------------------------.
* 1. Programmname: VarMZ_A.SPS (URL www.gesis.org/Dauerbeobachtung/ .
*       Mikrodaten/mikrodaten_tools/Varianz/VarMZ_A.SPS ) .
* 2. Programmautoren: Ulrich Rendtel (rendtel@em.uni-frankfurt.de) .
*       Bernhard Schimpl-Neimanns (schimpl-neimanns@zuma-mannheim.de) .
* 3. Zweck des Programms: Berechnung der Varianz des Regressionsschaetzers .
*       bei Randanpassung von Mikrozensus-Fallzahlen an die .
*       Bevoelkerungsfortschreibung im faktisch anonymisierten .
*       Mikrozensus (FAMZ) 1996.
*       Hier am Beispiel des Merkmals Y "Ledige" (EF35=1) fuer die .
*       Subpopulation Z "Bevoelkerung in Privathaushalten" (EF506=1) .
* 4. Weiterfuehrende Aufgabenbeschreibungen:  .
*       Rendtel, Ulrich; Schimpl-Neimanns, Bernhard: Varianzschaetzungen .
*          fuer den faktisch anonymisierten Mikrozensus. In: Jahrbuecher .
*          fuer Nationaloekonomie und Statistik, 220/6, 2000, S. 759-776.
*       Rendtel, Ulrich; Schimpl-Neimanns, Bernhard: Die Berechnung der  .
*          Varianz von Populationsschaetzern im Scientific Use File des .
*          Mikrozensus. In: ZUMA-Nachrichten Nr. 48, 2001, S. 85-116 .
*          (URL www.gesis.org/Publikationen/Zeitschriften/ .
*          ZUMA_Nachrichten/documents/pdfs/zn48_10-bernhard.pdf ) .
*       Schimpl-Neimanns, Bernhard; Rendtel, Ulrich: SAS-, SPSS- und .
*          STATA-Programme zur Berechnung der Varianz von  .
*          Populationsschaetzern im Mikrozensus. ZUMA-Methodenbericht .
*          Nr. 2001/04. Mannheim. (URL www.gesis.org/Publikationen/    .
*          Berichte/ZUMA_Methodenberichte/documents/pdfs/tb01_04.pdf ) .
* 5. Projektbeginn: Oktober 1999 .
* 6. Letzte Programmaenderung: 26. Januar 2001 .
* 7. Programmstatus: Getestet mit SPSS for Windows, Release 10.0.5, .
*       (Windows NT 4.0, SP6) und Mikrozensus 1996 (faktisch  .
*       anonymisierte 70%-Substichprobe).
* 8. Erforderliche Programmeingaben: SAS-Datensatz basierend auf den  .
*       Rohdaten des Mikrozensus ohne Missing Value Deklaration.
*       Schichtvariablen: EF1 Bundesland, EF708 Gemeindegroessenklasse, .
*          EF712 Gebaeudegroessenklasse .
*       Klumpenidentifikation (PSU): EF3 Auswahlbezirk .
*       Haushaltsidentifikation (HHNR): EF4 Haushaltsnummer
*       Abgrenzung der Anpassungsklassen (ANP): EF32 Geschlecht, .
*          EF52 Staatsangehoerigkeit, EF127 Stellung im Beruf .
*       Gruppendefinition (GRUPPE): EF1 Bundesland, ANP  .
*       Die bei anderen als in diesem Beispiel verwendeten Y-Variablen .
*       und Subpopulationen (Z) sowie insgesamt zu aendernden .
*       Programmschritte sind mit spitzen Klammern <> gekennzeichnet.
* 9. Grobe Programmstruktur:  .
*    Einzeldaten einlesen und benoetigte Variablen definieren .
*    Berechnen des Regressionskoeffizienten B^ .
*    Ausgabe des SOLL/IST Vergleichs fuer die Gruppen und des .
*       gewichteten Gesamt-Totals ueber die Gruppen  .
*    Berechnen der Hilfsgroesse u=g * (y - B^*1) .
*    Berechnen der Haushaltstotals der Hilfsgroesse  .
*    Haushaltsbezogene Daten weiterverarbeiten  .
*    Berechnen des Verhaeltnisses R=t_y / t_z .
*    Berechnen der Hilfsgroesse u=y - R*z .
*    Berechnen der PSU Totals und PSU Within Varianzen  .
*    Berechnen der Schicht Totals, der Between Varianz und der  .
*       Summe der gewichteten Within Varianzen .
*    Summation der Totals und Varianz Terme ueber die Schichten  .
*    Berechnung der Standard Abweichungen .
*    Berechnung der auszugebenden Kennwerte .
```

```
*    Ausgabe der Kennwerte .
*-----------------------------------------------------------------------.

set compression=on.
set header=no.
set mxwarns=300000.
set blanks=sysmis.
set width 80.
set length none.
set mxmemory=2097151.

* Teil A: Personen-/Individualdaten .

get file '<filename>'
   /keep ef1 ef3 ef4 ef32 <ef35> ef52 ef127 <ef506> ef708 ef712 <ef750>
   /rename (ef3 ef4 <ef750> = psu hhnr soll_ist).
weight off.
missing values all ().

* interessierendes Merkmal definieren .
compute y=0.
if (<ef35=1 & ef506=1>) y=1.
variable label y '<Fam.stand=ledig; Bev.in.Priv.HH (ef35=1 & ef506=1)>'.

* y_w:  mit Randanpassung gewichtete Beobachtung .
compute y_w=y*soll_ist.

* Protokollierung bis Ausgabe der Kennwerte abschalten.
* set printback none.

* Anpassungsklassen konstruieren.
compute anp=0.
if (ef52=1  & ef32=1 & ef127<>9 & ef127<>10) anp=1.
if (ef52=1  & ef32=2) anp=2.
if (ef52<>1 & ef32=1) anp=3.
if (ef52<>1 & ef32=2) anp=4.
if (ef52=1  & ef32=1 & ef127=9) anp=5.
if (ef52=1  & ef32=1 & ef127=10) anp=6.
variable label anp 'Anpassungsklasse'.
value label anp 1 'Deutsche Maenner'
                2 'Deutsche Frauen'
                3 'Auslaendische Maenner'
                4 'Auslaendische Frauen'
                5 'Zeit-/Berufssoldaten, BGS, Ber.pol.'
                6 'Grundwehrdienstleistende'.

* Gruppenvariable konstruieren.
compute gruppe=ef1*10+anp.
variable label gruppe 'Bundesland (EF1) * Anpassungsklasse'.

* Schichtvariable bilden: Bundesland (ef1) * .
*        Gemeindegroessenklasse (ef708) * Gebaeudeschicht (ef712).
compute schicht = ef1*100 + ef708*10 + ef712.
formats schicht (f4).
compute schicht=ef1*100+ef708*10+ef712.

* Sortieren nach Gruppen .
sort cases by gruppe.

* Personenfile (zwischen-) speichern .
save outfile '<mz_pers.sav>'
     /keep schicht psu hhnr gruppe y y_w soll_ist
     /compressed .

* Berechnen des Regressionskoeffizienten B_dach=t_y_g / t_x_g .
aggregate outfile=*
```

```
    /presorted
    /break gruppe
    /t_y_g = sum(y)
    /t_y_w = sum(y_w)
    /t_s_i = sum(soll_ist)
    /t_x_g = n
    /t_x_w = n.

compute B_dach=t_y_g/t_x_g.
compute t_y_w =t_y_w*100/0.7.
compute t_ist =t_x_w*100/0.7.
compute t_soll=t_ist *t_s_i/t_x_g.

* File mit Regressionskoeffizienten etc. (zwischen-) speichern.
save outfile '<reg_koef.sav>'.

* Berechnung des gewichteten Gesamt-Totals ueber die Gruppen .
match files file '<mz_pers.sav>'
          /table '<reg_koef.sav>'
          /by gruppe.

* Berechnen der Hilfsgroesse u=g*(y - B_dach*1) mit g=Soll_Ist .
compute u=Soll_Ist * (y-B_dach).

compute nw_stich=soll_ist*100/0.7.

* Sortieren nach Schicht, PSU und Haushalts-Nr.
sort cases by schicht psu hhnr.

* Berechnung der Haushaltstotals von u  .
aggregate outfile = *
         /presorted
         /break schicht psu hhnr
         /u_k = sum(u)
         /t_w = sum(y_w)
         /n_stichp 'Personen im Haushalt' = n
         /nw_stich=sum(nw_stich).

* Berechnen der PSU Totals und PSU Within Varianzen .
aggregate outfile = *
         /presorted
         /break schicht psu
         /psu_u = sum(u_k)
         /psu_var = sd(u_k)
         /psu_n = n
         /t_w = sum(t_w)
         /n_stichp 'Personen in PSU' = sum(n_stichp)
         /nw_stich=sum(nw_stich).

* Berechnen der Schicht Totals, der Between Varianz und der Summe der .
*           gewichteten Within Varianzen .
* SD(u_k) => VAR(u_k): .
compute psu_var=psu_var*psu_var.

* Einige PSUs sind im FAMZ nur mit E I N E M  Haushalt repraesentiert .
*        - Missings rekodieren .
recode psu_var (missing=0).

* Berechnung n_i*S^2(s_i)   .
compute N_Var=PSU_N*PSU_Var.

* Schicht Totals .
aggregate outfile = *
         /presorted
         /break schicht
         /str_u = sum(psu_u)
```

```
         /str_with = sum(n_var)
         /Between = sd(psu_u)
         /str_n = n
         /t_w = sum(t_w)
         /n_stichp 'Personen in Schicht' = sum(n_stichp)
         /nw_stich = sum(nw_stich).

* SD(psu_u) => VAR(psu_u): .
compute Between=Between*Between.
recode Between (missing=0).

* Strata Varianz = 100^2*0.99*n_(I,h)*S^2(n_I,h)/(0.7*0.7)  [= V_betw].
*                  + 100*0.3/(0.7*0.7) * Summe n_i*S^2(s_i) [= V_with].
compute V_betw=10000*0.99*STR_N*Between/(0.7*0.7).
compute V_with=  100*0.3 *STR_with/(0.7*0.7).
compute v=v_Betw + V_with.

* Summation der Varianz Terme ueber die Strata .
compute eins=1.
aggregate outfile = *
         /presorted
         /break eins
         /V_betw = sum(V_betw)
         /V_with = sum(V_with)
         /V = sum(V)
         /t_w = sum(t_w)
         /n_stichp 'Stichprobengroesse' = sum(n_stichp)
         /nw_stich = sum(nw_stich).

* Berechnung der Standardabweichungen .
compute sig_V=sqrt(V).
compute sig_B=sqrt(V_betw).
compute sig_W=sqrt(V_with).
compute t_w=t_w*100/0.7.
compute rel=sig_V/t_w.

* Ausgabe in 1000: .
compute t_w=t_w/1000.
compute nw_stich=nw_stich/1000.
compute sig_V=sig_V/1000.
compute sig_B=sig_B/1000.
compute sig_W=sig_W/1000.
* Ausgabe in Prozent:.
compute rel=rel*100.

variable label t_w 'Hochgerechneter Merkmalswert (in 1000)'.
variable label nw_stich 'hochger. Stichprobengroesse'.
variable label sig_V 'Std.Fehler Merkmalswert (in 1000)'.
variable label sig_B 'Std.Fehler Merkmalswert Between-Teil (in 1000)'.
variable label sig_W 'Std.Fehler Merkmalswert Within-Teil (in 1000)'.
variable label rel 'relativer Std.Fehler (sig_V*100/t_w)'.

formats n_stichp nw_stich (f6)
        t_w (f7.1)
        sig_v sig_b sig_w (f10.4)
        rel (f5.4).

* Ausgabe der Kennwerte.
set printback both.
SUMMARIZE
  /TABLES= t_w sig_v sig_b sig_w rel
  /FORMAT=NOLIST /CELLS=FIRST .
```

```
/* ------------------------- VarMZ_T.DO --------------------------------
   1. Programmname: VarMZ_T.DO (URL www.gesis.org/Dauerbeobachtung/
         Mikrodaten/mikrodaten_tools/Varianz/VarMZ_T.DO )
   2. Programmautoren: Ulrich Rendtel (rendtel@em.uni-frankfurt.de)
         Bernhard Schimpl-Neimanns (schimpl-neimanns@zuma-mannheim.de)
   3. Zweck des Programms: Berechnung der Varianz des Pi-Schaetzers fuer
         ein Merkmal Y im faktisch anonymisierten Mikrozensus (FAMZ) 1996.
         Hier am Beispiel des Merkmals Y "Ledige" (EF35=1) fuer die
         Subpopulation Z "Bevoelkerung in Privathaushalten" (EF506=1)
   4. Weiterfuehrende Aufgabenbeschreibungen:
         Rendtel, Ulrich; Schimpl-Neimanns, Bernhard: Varianzschaetzungen
            fuer den faktisch anonymisierten Mikrozensus. In: Jahrbuecher
            fuer Nationaloekonomie und Statistik, 220/6, 2000, S. 759-776.
         Rendtel, Ulrich; Schimpl-Neimanns, Bernhard: Die Berechnung der
            Varianz von Populationsschaetzern im Scientific Use File des
            Mikrozensus. In: ZUMA-Nachrichten Nr. 48, 2001, S. 85-116.
           (URL www.gesis.org/Publikationen/Zeitschriften/
           ZUMA_Nachrichten/documents/pdfs/zn48_10-bernhard.pdf )
         Schimpl-Neimanns, Bernhard; Rendtel, Ulrich: SAS-, SPSS- und
            STATA-Programme zur Berechnung der Varianz von
            Populationsschaetzern im Mikrozensus. ZUMA-Methodenbericht
             Nr. 2001/04. Mannheim. (URL www.gesis.org/Publikationen/
             Berichte/ZUMA_Methodenberichte/documents/pdfs/tb01_04.pdf )
   5. Projektbeginn: Oktober 1999
   6. Letzte Programmaenderung: 26. Januar 2001
   7. Programmstatus: Getestet mit Intercooled STATA 6.0 for Windows
         95/98/NT (Windows NT 4.0, SP6) und Mikrozensus 1996 (faktisch
         anonymisierte 70%-Substichprobe)
   8. Erforderliche Programmeingaben: STATA-Datensatz basierend auf den
         Rohdaten des Mikrozensus. Das File sollte keine Missing Values
         enthalten.
         Schichtvariablen: EF1 Bundesland, EF708 Gemeindegroessenklasse,
            EF712 Gebaeudegroessenklasse
         Klumpenidentifikation (PSU): EF3 Auswahlbezirk
         Haushaltsidentifikation (HHNR): EF4 Haushaltsnummer
         Die bei anderen als in diesem Beispiel verwendeten Y-Variablen
         und Subpopulationen (Z) sowie insgesamt zu aendernden
         Programmschritte sind mit spitzen Klammern <> gekennzeichnet.
   9. Grobe Programmstruktur:
         A Einzeldaten einlesen und benoetigte Variablen definieren
           Berechnen der Haushaltstotals
         B Haushaltsbezogene Daten weiterverarbeiten (ggf. einlesen)
         C Berechnen der PSU Totals und PSU Within Varianzen
           Berechnen der Schicht Totals, der Between Varianz und der
              Summe der gewichteten Within Varianzen
           Summation der Totals und Varianz Terme ueber die Schichten
           Berechnung der Standardabweichungen
           Berechnung der Varianz unter Annahme der Binomialverteilung
           Berechnung der auszugebenden Kennwerte
           Ausgabe der Kennwerte
---------------------------------------------------------------------- */

version 6.0
log using <varmz_t.log>, replace
set log linesize 250
set memory <80m>
* set virtual on
set more off
```

```
#delimit ;

   /* Teil A: Daten Personenbezogen  */

use ef1 ef3 ef4 <ef35> <ef506> ef708 ef712 using <filename>, clear ;
rename ef3 psu ;
rename ef4 hhnr ;

generate schicht = ef1*100 + ef708*10 + ef712;
generate y = (ef35==1)*(ef506==1) ;
generate z = (ef506==1) ;

label variable psu "Auswahlbezirksnummer (EF3)" ;
label variable hhnr "lfd. Haushaltsnummer (EF4)" ;
label variable schicht "Bu.land | Gem.groesse | Geb.schicht" ;
label variable y "Ledige in Privathaushalten (1, sonst 0)" ;
label variable z "Bevoelkerung in Privathaushalten (1, sonst 0)" ;

   /* Sortieren falls noetig, sonst ueberspringen */
sort schicht psu hhnr;

   /* Berechnen der Haushaltstotals */
collapse (sum) y_k=y z_k=z, by(schicht psu hhnr) ;

   /* Ende Teil A */

   /* Teil B: Daten Haushaltsbezogen
                   Beim Einlesen haushaltsbezogener Daten muessen
                   Y und Z als y_k und z_k aggregiert vorliegen */

    /* Sortieren falls noetig, sonst ueberspringen */
sort schicht psu hhnr;

    /* Ende Teil B */


    /* Teil C: Ab hier weiter in beiden Faellen  */

    /* Berechnen der PSU Totals und PSU Within Varianzen */
collapse (sum) psu_y=y_k n_stichp=z_k
         (count) psu_n=y_k
         (sd) psu_var=y_k,
       by(schicht psu) ;
* psu_var ist nur Std.Abw. (sd): quadrieren ;
replace psu_var = psu_var*psu_var ;

    /* Berechnen der Schicht Totals, der Between Varianz und der
       Summe der gewichteten Within Varianzen */

    /* Einige PSUs sind im FAMZ nur mit E I N E M  Haushalt
           repraesentiert: Missings rekodieren */
recode psu_var .=0 ;
        /* Berechnung n_i*S^2(s_i)   */
generate n_var = psu_n * psu_var ;
label var n_var "PSU Within Varianz" ;

*    AGGREGATION PSU- => Schicht-EBENE ;
collapse (sum) str_y=psu_y str_with=n_var n_stichp
         (count) str_n=psu_y
```

```
        (sd) between=psu_y,
       by(schicht);
*   Missing Vaules rekodieren ;
recode between .=0 ;
*   Std.Abw. => Varianz ;
replace between=between*between ;

  /* Strata Varianz = 100^2*0.99*n_(I,h)*S^2(n_I,h)          [= V_betw ]
                     +100*0.3/(0.7*0.7) * Summe n_i*S^2(s_i) [= V_with ] */
generate total=100*str_y/0.7 ;
generate v_betw=100*100*0.99*str_n*between/(0.7*0.7) ;
generate v_with=100*0.3*str_with/(0.7*0.7) ;
generate v = v_betw+v_with ;

   /* Summation der Totals und Varianz Terme ueber die Strata */
collapse (sum) total v v_betw v_with n_stichp ;

   /* Berechnung der Standard Abweichungen */
generate sig_v = sqrt(v) ;
generate sig_b = sqrt(v_betw) ;
generate sig_w = sqrt(v_with) ;
generate rel = sig_v / total ;

   /* Berechnung der Binomial Varianz
      hier: Y-Merkmal und Anzahl der zur Subpopulation gehoerenden
      Personen liegen aggregiert vor (y_k, n_stichp)
        --- bei total Hochrechnung (100/0.7) zuruecknehmen --- */
generate p_dach = (total*(0.7/100))/n_stichp ;
generate rel_bin = sqrt( 0.99*(1-p_dach)/(p_dach*(n_stichp-1)) ) ;
generate deft=rel/rel_bin;
* Ausgabe in 1000 ;
replace total = total/1000;
replace sig_v = sig_v/1000;
replace sig_b = sig_b/1000;
replace sig_w = sig_w/1000;
* Ausgabe in Prozent ;
replace rel = rel*100;
replace rel_bin = rel_bin*100;
replace p_dach=p_dach*100;

label variable total   "Total (in 1000)" ;
label variable sig_v   "Std.Fehler (in 1000)" ;
label variable sig_b   "Std.Fehler Between-Teil (in 1000)" ;
label variable sig_w   "Std.Fehler Within-Teil (in 1000)" ;
label variable rel     "relativer Std.Fehler (in %)" ;
label variable deft    "Design-Effekt Faktor des Std.Fehlers" ;
label variable rel_bin "Relativer Std.Fehler Binomialverteilung (in %)" ;
label variable p_dach  "Anteil Y in Subpopulation Z (in %)";

list total sig_v sig_b sig_w rel deft rel_bin p_dach ;

#delimit cr
log close
exit
```

```
/* ------------------------- VarMZ_R.DO --------------------------------
   1. Programmname: VarMZ_R.DO (URL www.gesis.org/Dauerbeobachtung/
         Mikrodaten/mikrodaten_tools/Varianz/VarMZ_R.DO )
   2. Programmautoren: Ulrich Rendtel (rendtel@em.uni-frankfurt.de)
         Bernhard Schimpl-Neimanns (schimpl-neimanns@zuma-mannheim.de)
   3. Zweck des Programms: Berechnung der Varianz des Pi-Schaetzers fuer
         das Verhaeltnis R zweier Totals, t_y und t_z, im faktisch
         anonymisierten Mikrozensus (FAMZ) 1996.
         Hier am Beispiel des Merkmals Y "Ledige" (EF35=1) fuer die
         Subpopulation Z "Bevoelkerung in Privathaushalten" (EF506=1)
         Das Programm kann auch zur Berechnung der Varianz des
         arithmetischen Mittelwerts der Variablen Y fuer die Subpopulation
         Z verwendet werden. (Ein Beispiel befindet sich am Ende des
         Programms.)
   4. Weiterfuehrende Aufgabenbeschreibungen:
         Rendtel, Ulrich; Schimpl-Neimanns, Bernhard: Varianzschaetzungen
            fuer den faktisch anonymisierten Mikrozensus. In: Jahrbuecher
            fuer Nationaloekonomie und Statistik, 220/6, 2000, S. 759-776.
         Rendtel, Ulrich; Schimpl-Neimanns, Bernhard: Die Berechnung der
            Varianz von Populationsschaetzern im Scientific Use File des
            Mikrozensus. In: ZUMA-Nachrichten Nr. 48, 2001, S. 85-116.
           (URL www.gesis.org/Publikationen/Zeitschriften/
           ZUMA_Nachrichten/documents/pdfs/zn48_10-bernhard.pdf )
         Schimpl-Neimanns, Bernhard; Rendtel, Ulrich: SAS-, SPSS- und
            STATA-Programme zur Berechnung der Varianz von
            Populationsschaetzern im Mikrozensus. ZUMA-Methodenbericht
            Nr. 2001/04. Mannheim. (URL www.gesis.org/Publikationen/
            Berichte/ZUMA_Methodenberichte/documents/pdfs/tb01_04.pdf )
   5. Projektbeginn: Oktober 1999
   6. Letzte Programmaenderung: 26. Januar 2001
   7. Programmstatus: Getestet mit Intercooled STATA 6.0 for Windows
        95/98/NT (Windows NT 4.0, SP6) und Mikrozensus 1996 (faktisch
        anonymisierte 70%-Substichprobe).
   8. Erforderliche Programmeingaben: STATA-Datensatz basierend auf den
         Rohdaten des Mikrozensus. Das File sollte keine Missing Values
         enthalten.
         Schichtvariablen: EF1 Bundesland, EF708 Gemeindegroessenklasse,
            EF712 Gebaeudegroessenklasse
         Klumpenidentifikation (PSU): EF3 Auswahlbezirk
         Haushaltsidentifikation (HHNR): EF4 Haushaltsnummer
         Die bei anderen als in diesem Beispiel verwendeten Y-Variablen
         und Subpopulationen (Z) sowie insgesamt zu aendernden
         Programmschritte sind mit spitzen Klammern <> gekennzeichnet.
   9. Grobe Programmstruktur:
      A Einzeldaten einlesen und benoetigte Variablen definieren
        Berechnen des Gesamt-Totals, des Verhaeltnisses R = t_y / t_z
           und der Hilfsgroesse u = y-R*z
        Berechnen der Haushaltstotals
      B Haushaltsbezogene Daten weiterverarbeiten (ggf. einlesen)
      C Berechnen der PSU Totals und PSU Within Varianzen
        Berechnen der Schicht Totals, der Between Varianz und der
           Summe der gewichteten Within Varianzen
        Summation der Totals und Varianz Terme ueber die Schichten
        Berechnung der Standard Abweichungen
        Berechnung der Varianz unter Annahme der Binomialverteilung
        Berechnung der auszugebenden Kennwerte
        Ausgabe der Kennwerte
   ---------------------------------------------------------------- */
```

```
version 6.0
log using <varmz_r.log>, replace
set log linesize 250
set memory <80m>
* set virtual on
set more off
#delimit ;

   /* Teil A:  Daten Personenbezogen  */

use ef1 ef3 ef4 <ef35> <ef506> ef708 ef712 using <filename>, clear ;
rename ef3 psu ;
rename ef4 hhnr ;

generate schicht = ef1*100 + ef708*10 + ef712;
generate z = <(ef506==1)> ;
generate y = <(ef35==1)>*z ;

label variable psu "Auswahlbezirksnummer (EF3)" ;
label variable hhnr "lfd. Haushaltsnummer (EF4)" ;
label variable schicht "Bundesland | Gemeindegroesse | Gebaeudeschicht" ;
label variable y "<Ledige in Privathaushalten (1, sonst 0)>" ;
label variable z "<Bevoelkerung in Privathaushalten (1, sonst 0)>" ;

   /* Sortieren falls noetig, sonst ueberspringen */
sort schicht psu hhnr;

* Berechnen der Gesamt-Totals (t_y, t_z), des Verhaeltnisses ;
*    und der Hilfsgroesse u = y - R*z  ("ratio residual") ;
egen t_y = sum(y) ;
egen t_z = sum(z) ;
generate r = t_y / t_z ;
generate u = y - r*z ;
generate n_stichp=t_z ;
label variable t_y "Total y" ;
label variable t_z "Total z" ;
label variable r "Verhaeltnis R = t_y / t_z" ;
label variable u "Hilfsgroesse y-R*z" ;

   /* Berechnen der Haushaltstotals und Uebernahme der Gesamt-Totals
         sowie des Verhaeltnisses R  */
collapse (sum) u_k=u z_k=z
         (max) n_stichp r t_y t_z,
       by(schicht psu hhnr);
recode u_k .=0 ;

   /* Ende Teil A */


   /* Teil B: Daten Haushaltsbezogen
                   Beim Einlesen haushaltsbezogener Daten muessen
                   Y, Z und U etc. als y_k, z_k und u_k aggregiert
                   vorliegen (siehe oben) */

    /* Sortieren falls noetig, sonst ueberspringen */
sort schicht psu hhnr;

    /* Ende Teil B */
```

```
    /* Teil C: Ab hier weiter in beiden Faellen  */

   /* Berechnen der PSU Totals und PSU Within Varianzen */
collapse (sum) psu_u=u_k
         (count) psu_n=u_k
         (sd) psu_var=u_k
         (max) n_stichp r t_y t_z,
       by(schicht psu) ;
recode psu_u .=0 ;
* psu_var ist Std.Abw.: quadrieren;
replace psu_var = psu_var*psu_var ;

    /* Berechnen der Schicht Totals, der Between Varianz und der Summe der
       gewichteten Within Varianzen */
         /* Einige PSUs sind im FAMZ nur mit E I N E M  Haushalt
            repraesentiert - Missings rekodieren */
recode psu_var .=0 ;
*    Berechnung n_i*S^2(s_i);
generate n_var = psu_n * psu_var ;
label variable n_var "PSU Within Varianz" ;

collapse (sum) str_u=psu_u str_with=n_var
         (count) str_n=psu_u
         (sd) between=psu_u
         (max) n_stichp r t_y t_z,
       by(schicht) ;
recode between .=0 ;
* between ist std.abw. => quadrieren ;
replace between=between*between ;

  /* Strata Varianz = 100^2*0.99*n_(I,h)*S^2(n_I,h)/(0.7*0.7) + [= V_betw]
                      100*0.3/(0.7*0.7) * Summe n_i*S^2(s_i)    [= V_with]*/
generate v_betw=100*100*0.99*str_n*between/(0.7*0.7) ;
generate v_with=100*0.3*str_with/(0.7*0.7) ;
generate v = v_betw+v_with ;

   /* Summation der Varianz Terme ueber die Strata */
collapse (sum) v v_betw v_with
         (max) n_stichp r t_y t_z ;

* Stichproben- -> Populationswerte ;
replace t_y=t_y*100/0.7 ;
replace t_z=t_z*100/0.7 ;

   /* Berechnung der Standardabweichungen und Ausdrucken der Ergebnisse
        Division der Varianzen durch t_z^2 */
replace v=v/(t_z*t_z) ;
replace v_betw=v_betw/(t_z*t_z) ;
replace v_with=v_with/(t_z*t_z) ;
generate sig_v = 100*sqrt(v) ;
generate sig_b = 100*sqrt(v_betw) ;
generate sig_w = 100*sqrt(v_with) ;
generate rel = sig_v / r ;

* Berechnung des Std. Fehlers bei einfacher Zufallsstichprobe ;
*    NUR BEI ANTEILSWERTEN ;
```

```
generate sig_bin=100*sqrt( (r*(1-r))/(n_stichp-1));
generate deft=sig_v/sig_bin;

label variable sig_v "Std.Fehler Merkmalswert (in Prozent)" ;
label variable sig_b "Std.Fehler Between-Teil (in Prozent)" ;
label variable sig_w "Std.Fehler Within-Teil (in Prozent)" ;
label variable sig_bin "Std.Fehler Binomialverteilung (in Prozent)";
label variable rel "relativer Std.Fehler in Prozent (sig_v/r)";
label variable deft "Design-Effekt Faktor";
label variable r "Anteilswert R (in Prozent)";
label variable t_y "Total y (in 1000)";
label variable t_z "Total z (in 1000)";

* Ausgabe der Kennwerte ;
* t_y und t_z in 1000 und r in Prozent ausgeben ;
replace t_y=t_y / 1000;
replace t_z=t_z / 1000 ;
replace r=r*100;
format t_y t_z r sig_v sig_b sig_w rel %10.6f ;
list t_y t_z r ;
list sig_v sig_b sig_w rel ;
list deft sig_bin ;

#delimit cr
log close
exit

/* =============== PROGRAMMAENDERUNGEN BEI MITTELWERTEN ====================
   Beispiel: Merkmal Y "Heiratsalter von verheirateten und mit dem Partner
                        zusammenlebenden Frauen ..."
             Subpopulation Z "verheiratete, mit dem Partner zusammenlebende
                        Frauen, Bevoelkerung am Hauptwohnsitz, gueltige
                        Angaben zum Heiratsjahr und Heiratsjahr>=1925,
                        Geburtsjahr>=1901" .

generate z =<(ef32==2 & ef505>=1 & ef505<=2 & ef35==2 & ef575>=1 & ef575<=3
             & ef33>=1901 & ef36>=1925 & ef36<=1996)> ;
generate y =<z*(ef36-ef33)>;
label variable y "<Heiratsalter v. Frauen ...>";
label variable z "<verh. Frauen, Bevoelkerung am Hauptwohns. (1, sonst 0)>" ;

* (...) ;
   /* Berechnung der Standardabweichungen und Ausdrucken der Ergebnisse */
* Ausgabe des Mittelwerts r nicht in Prozent: ;
*         naechste Zeile loeschen oder auskommentieren ;
* replace r=r*100;
label variable r "Mittelwert von y";

================== PROGRAMMAENDERUNGEN BEI MITTELWERTEN ====================*/
```

```
/* ------------------------- VarMZ_A.DO ---------------------------------
   1. Programmname: VarMZ_A.DO (URL www.gesis.org/Dauerbeobachtung/
         Mikrodaten/mikrodaten_tools/Varianz/VarMZ_A.DO )
   2. Programmautoren: Ulrich Rendtel (rendtel@em.uni-frankfurt.de)
         Bernhard Schimpl-Neimanns (schimpl-neimanns@zuma-mannheim.de)
   3. Zweck des Programms: Berechnung der Varianz des Regressionsschaetzers
         bei Randanpassung von Mikrozensus-Fallzahlen an die
         Bevoelkerungsfortschreibung im faktisch anonymisierten
         Mikrozensus (FAMZ) 1996.
         Hier am Beispiel des Merkmals Y "Ledige" (EF35=1) fuer die
         Subpopulation Z "Bevoelkerung in Privathaushalten" (EF506=1)
   4. Weiterfuehrende Aufgabenbeschreibungen:
         Rendtel, Ulrich; Schimpl-Neimanns, Bernhard: Varianzschaetzungen
            fuer den faktisch anonymisierten Mikrozensus. In: Jahrbuecher
            fuer Nationaloekonomie und Statistik, 220/6, 2000, S. 759-776.
         Rendtel, Ulrich; Schimpl-Neimanns, Bernhard: Die Berechnung der
            Varianz von Populationsschaetzern im Scientific Use File des
            Mikrozensus. In: ZUMA-Nachrichten Nr. 48, 2001,S. 85-116.
            (URL www.gesis.org/Publikationen/Zeitschriften/
            ZUMA_Nachrichten/documents/pdfs/zn48_10-bernhard.pdf )
         Schimpl-Neimanns, Bernhard; Rendtel, Ulrich: SAS-, SPSS- und
            STATA-Programme zur Berechnung der Varianz von
            Populationsschaetzern im Mikrozensus. ZUMA-Methodenbericht
            Nr. 2001/04. Mannheim. (URL www.gesis.org/Publikationen/
            Berichte/ZUMA_Methodenberichte/documents/pdfs/tb01_04.pdf )
   5. Projektbeginn: Oktober 1999
   6. Letzte Programmaenderung: 26. Januar 2001
   7. Programmstatus: Getestet mit Intercooled STATA 6.0 for Windows
         95/98/NT (Windows NT 4.0, SP6) und Mikrozensus 1996 (faktisch
         anonymisierte 70%-Substichprobe).
   8. Erforderliche Programmeingaben: STATA-Datensatz basierend auf den
         Rohdaten des Mikrozensus. Das File sollte keine Missing Values
         enthalten.
         Schichtvariablen: EF1 Bundesland, EF708 Gemeindegroessenklasse,
            EF712 Gebaeudegroessenklasse
         Klumpenidentifikation (PSU): EF3 Auswahlbezirk
         Haushaltsidentifikation (HHNR): EF4 Haushaltsnummer
         Abgrenzung der Anpassungsklassen (ANP): EF32 Geschlecht,
            EF52 Staatsangehoerigkeit, EF127 Stellung im Beruf
         Gruppendefinition (GRUPPE): EF1 Bundesland, ANP
         Die bei anderen als in diesem Beispiel verwendeten Y-Variablen
         und Subpopulationen (Z) sowie insgesamt zu aendernden
         Programmschritte sind mit spitzen Klammern <> gekennzeichnet.
   9. Grobe Programmstruktur:
       Einzeldaten einlesen und benoetigte Variablen definieren
       Berechnen des Regressionskoeffizienten B^
       Ausgabe des SOLL/IST Vergleichs fuer die Gruppen und des
          gewichteten Gesamt-Totals ueber die Gruppen
       Berechnen der Hilfsgroesse u=g * (y - B^*1)
       Berechnen der Haushaltstotals der Hilfsgroesse
       Haushaltsbezogene Daten weiterverarbeiten
       Berechnen des Verhaeltnisses R=t_y / t_z
       Berechnen der Hilfsgroesse u=y - R*z
       Berechnen der PSU Totals und PSU Within Varianzen
       Berechnen der Schicht Totals, der Between Varianz und der
          Summe der gewichteten Within Varianzen
       Summation der Totals und Varianz Terme ueber die Schichten
       Berechnung der Standard Abweichungen
```

```
        Berechnung der auszugebenden Kennwerte
        Ausgabe der Kennwerte
   ------------------------------------------------------------------- */

veersion 6.0
log using <varmz_a.log>, replace
set log linesize 250
set memory <80m>
* set virtual on
set more off
#delimit ;

   /* Teil A:  Daten Personenbezogen  */

use ef1 ef3 ef4 ef708 ef712 ef32 ef52 ef127 <ef35> <ef506> <ef750>
    using <filename>, clear ;
rename ef3 psu ;
rename ef4 hhnr ;
rename <ef750> soll_ist ;

generate schicht = ef1*100 + ef708*10 + ef712;
generate z = <(ef506==1)>;
generate y =<z*(ef35==1)>;
generate anp= (1*(ef32==1 & ef52==1 & ef127~=9 & ef127~=10)) +
              (2*(ef32==2 & ef52==1)) +
              (3*(ef32==1 & ef52~=1)) +
              (4*(ef32==2 & ef52~=1)) +
              (5*(ef32==1 & ef52==1 & ef127==9)) +
              (6*(ef32==1 & ef52==1 & ef127==10));
* anp: 1=Deutsche Maenner, 2=Deutsche Frauen, 3=Auslaendische Maenner,
       4=Auslaendische Frauen, 5=Zeit-/Berufssoldaten, 6=Wehrpflichtige ;
generate gruppe=ef1*10+anp;
gen y_w=y*soll_ist;
label variable psu "Auswahlbezirksnummer (EF3)" ;
label variable hhnr "lfd. Haushaltsnummer (EF4)" ;
label variable schicht "Bundesland | Gemeindegroesse | Gebaeudeschicht" ;
label variable y "<Ledige in Privathaushalten (1, sonst 0)>";
label variable z "<Bevoelkerung in Privathaushalten (1, sonst 0)>" ;
label variable anp "Anpassungsklassen"
label variable gruppe "Hochrechnungsgruppen";
label variable soll_ist "<Personen-Hochrechnungsfaktor (EF750)>";
label variable y_w "Merkmal Y * soll_ist - gewichtete Beobachtung";

  /* Sortieren nach Gruppen */
sort gruppe ;

  /* Berechnen des Regressionskoeffizienten B_dach=t_y_g / t_x_g  */
egen t_y_g=sum(y), by(gruppe);
egen t_y_w=sum(y_w), by(gruppe);
egen t_s_i=sum(soll_ist), by(gruppe);
egen t_x_g=count(y), by(gruppe);
egen t_x_w=count(y_w), by(gruppe);
generate B_dach =t_y_g/t_x_g ;
replace t_y_w=t_y_w*100/0.7 ;
generate t_ist =t_x_w*100/0.7 ;
generate t_soll=t_ist *t_s_i/t_x_g ;

   /* Falls Ausgabe des Soll/Ist Vergleichs fuer die einzelnen Gruppen
```

```
      benoetigt: "*" loeschen */
* sort gruppe ;
* by gruppe: summarize B_dach t_ist t_soll t_y_w ;

   /* Berechnung des gewichteten Gesamt-Totals ueber die Gruppen */
egen t_w=sum(y_w) ;
replace t_w=t_w*100/0.7 ;
label variable t_w "gewichtetes Gesamt-Total" ;

   /* Berechnen der Hilfsgroesse u=g*(y - B_dach*1) mit g=Soll_Ist  */
generate u=soll_ist*(y-B_dach) ;

  /* Sortieren nach Schicht PSU und Haushalt  */
sort schicht psu hhnr ;

* Berechnung der Haushaltstotals von u ;
generate n_stichp=_N ;
label variable n_stichp "Stichprobengroesse n" ;
collapse (sum) u_k=u
         (max) n_stichp t_w,
       by(schicht psu hhnr) ;
recode u_k .=0 ;

   /* Berechnen der PSU Totals und PSU Within Varianzen  */
collapse (sum) psu_u=u_k
         (count) psu_n=u_k
         (sd) psu_var=u_k
         (max) n_stichp t_w,
       by(schicht psu);
recode psu_u .=0 ;
* psu_var ist nur Std.Abw.: quadrieren ;
replace psu_var = psu_var*psu_var ;

   /* Berechnen der Schicht Totals, der Between Varianz und der Summe der
      gewichteten Within Varianzen
          Einige PSUs sind im FAMZ nur mit E I N E M  Haushalt repraesentiert
          - Missings rekodieren */
recode psu_var .=0 ;
*    Berechnung n_i*S^2(s_i)  ;
generate n_var = psu_n * psu_var ;
label variable n_var "PSU Within Varianz" ;

collapse (sum) str_u=psu_u str_with=n_var
         (count) str_n=psu_u
         (sd) between=psu_u
         (max) n_stichp t_w,
       by(schicht);
recode between .=0 ;
replace between=between*between;

   /* Strata Varianz=100*100*0.99*n_(I,h)*S^2(n_I,h)/(0.7*0.7) + [= V_betw]
*                    100*0.3/(0.7*0.7) * Summe n_i*S^2(s_i)      [= V_with] */
generate v_betw=100*100*0.99*str_n*between/(0.7*0.7);
generate v_with=100*0.3*str_with/(0.7*0.7);
generate v = v_betw+v_with;

   /* Summation der Varianz Terme ueber die Strata */
collapse (sum) v v_betw v_with
```

```
         (max) n_stichp t_w ;

   /* Berechnung der Standard Abweichungen und Ausdrucken der Ergebnisse */
generate sig_v = sqrt(v);
generate sig_b = sqrt(v_betw);
generate sig_w = sqrt(v_with);
generate rel = sig_v / t_w;

* Ausgabe in 1000 ;
replace t_w = t_w/1000 ;
replace sig_v = sig_v/1000;
replace sig_b = sig_b/1000;
replace sig_w = sig_w/1000;
* Ausgabe in Prozent ;
replace rel = rel*100 ;

label variable t_w "gewichtetes/angepasstes Total (in 1000)" ;
label variable sig_v "Std.Fehler Merkmalswert (in 1000)";
label variable sig_b "Std.Fehler Between-Teil (in 1000)";
label variable sig_w "Std.Fehler Within-Teil (in 1000)";
label variable rel "relativer Std.Fehler (sig_v/r) (in Prozent)";

list t_w sig_v sig_b sig_w rel ;

#delimit cr

log close
exit
```